The Eukaryotic Chromatin Computer

Components, Mode of Action, Properties, Tasks, Computational Power, and Disease Relevance

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

Vorgelegt von

Diplom-Informatiker Christian Arnold,
geboren am 22. Juli 1984 in Borna

Die Annahme der Dissertation wurde empfohlen von:
1. Jun.-Prof. Dr. Sonja Prohaska, Universität Leipzig
2. Prof. Dr. Manfred D. Laubichler, Arizona State University

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 14.02.2014 mit dem Gesamtprädikat '*summa cum laude*'.

Abstract

Eukaryotic genomes are typically organized as chromatin, the complex of DNA and proteins that forms chromosomes within the cell's nucleus. Chromatin has pivotal roles for a multitude of functions, most of which are carried out by a complex system of covalent chemical modifications of histone proteins.

The propagation of patterns of these histone post-translational modifications across cell divisions is particularly important for maintenance of the cell state in general and the transcriptional program in particular. The discovery of epigenetic inheritance phenomena — mitotically and/or meiotically heritable changes in gene function resulting from changes in a chromosome without alterations in the DNA sequence — was remarkable because it disproved the assumption that information is passed to daughter cells exclusively through DNA. However, DNA replication constitutes a dramatic disruption of the chromatin state that effectively amounts to partial erasure of stored information. To preserve its epigenetic state the cell reconstructs (at least part of) the histone post-translational modifications by means of processes that are still very poorly understood. A plausible hypothesis is that the different combinations of reader and writer domains in histone-modifying enzymes implement local rewriting rules that are capable of "recomputing" the desired parental patterns of histone post-translational modifications on the basis of the partial information contained in that half of the nucleosomes that predate replication.

It is becoming increasingly clear that both information processing and computation are omnipresent and of fundamental importance in many fields of the natural sciences and the cell in particular. The latter is exemplified by the increasingly popular research areas that focus on computing with DNA and membranes. Recent work suggests that during evolution, chromatin has been converted into a powerful cellular memory device capable of storing and processing large amounts of information. Eukaryotic chromatin may therefore also act as a cellular computational device capable of performing actual computations in a biological context. A recent theoretical study indeed demonstrated that even relatively simple models of chromatin computation are computationally universal and hence conceptually more powerful than gene regulatory networks.

In the first part of this thesis, I establish a deeper understanding of the computational capacities and limits of chromatin, which have remained largely unexplored. I analyze selected biological building blocks of the chromatin computer and compare it to system components of general purpose computers, particularly focusing on memory and the logical and arithmetical operations. I argue that it has a massively parallel architecture, a set of read-write rules that operate non-deterministically on chromatin, the capability of self-modification, and more generally striking analogies to amorphous computing. I therefore propose a cellular automata-like 1-D string as its computational paradigm on which sets of local rewriting rules are applied asynchronously with time-dependent probabilities. Its mode of operation is therefore conceptually similar to well-known concepts from the complex systems theory. Furthermore, the chromatin computer provides volatile memory with a massive information content that can be exploited by the cell. I estimate that its memory size lies in the realms of several hundred megabytes of writable information per cell, a value that I compare

with DNA itself and *cis*-regulatory modules. I furthermore show that it has the potential to not only perform computations in a biological context but also in a strict informatics sense. At least theoretically it may therefore be used to calculate any computable function or algorithm more generally. Chromatin is therefore another representative of the growing number of non-standard computing examples.

As an example for a biological challenge that may be solved by the "chromatin computer", I formulate epigenetic inheritance as a computational problem and develop a flexible stochastic simulation system for the study of recomputation-based epigenetic inheritance of individual histone post-translational modifications. The implementation uses Gillespie's stochastic simulation algorithm for exactly simulating the time evolution of the chemical master equation of the underlying stochastic process. Furthermore, it is efficient enough to use an evolutionary algorithm to find a system of enzymes that can stably maintain a particular chromatin state across multiple cell divisions. I find that it is easy to evolve such a system of enzymes even without explicit boundary elements separating differentially modified chromatin domains. However, the success of this task depends on several previously unanticipated factors such as the length of the initial state, the specific pattern that should be maintained, the time between replications, and various chemical parameters. All these factors also influence the accumulation of errors in the wake of cell divisions.

Chromatin-regulatory processes and epigenetic (inheritance) mechanisms constitute an intricate and sensitive system, and any misregulation may contribute significantly to various diseases such as Alzheimer's disease. Intriguingly, the role of epigenetics and chromatin-based processes as well as non-coding RNAs in the etiology of Alzheimer's disease is increasingly being recognized. In the second part of this thesis, I explicitly and systematically address the two hypotheses that (i) a dysregulated chromatin computer plays important roles in Alzheimer's disease and (ii) Alzheimer's disease may be considered as an evolutionarily young disease. In summary, I found support for both hypotheses although for hypothesis 1, it is very difficult to establish causalities due to the complexity of the disease. However, I identify numerous chromatin-associated, differentially expressed loci for histone proteins, chromatin-modifying enzymes or integral parts thereof, non-coding RNAs with guiding functions for chromatin-modifying complexes, and proteins that directly or indirectly influence epigenetic stability (e.g., by altering cell cycle regulation and therefore potentially also the stability of epigenetic states).

For the identification of differentially expressed loci in Alzheimer's disease, I use a custom expression microarray that was constructed with a novel bioinformatics pipeline. Despite the emergence of more advanced high-throughput methods such as RNA-seq, microarrays still offer some advantages and will remain a useful and accurate tool for transcriptome profiling and expression studies. However, it is non-trivial to establish an appropriate probe design strategy for custom expression microarrays because alternative splicing and transcription from non-coding regions are much more pervasive than previously appreciated. To obtain an accurate and complete expression atlas of genomic loci of interest in the post-ENCODE era, this additional transcriptional complexity must be considered during microarray design and requires well-considered probe design strategies that are often neglected. This encompasses, for example, adequate preparation of a set of target sequences and accurate estimation of probe specificity. With the help of this pipeline, two custom-tailored microarrays have been constructed that include a comprehensive collection of non-coding RNAs. Additionally, a user-friendly web server has been set up that makes the developed pipeline publicly available for other researchers.

Zusammenfassung

Eukaryotische Genome sind typischerweise in Form von Chromatin organisiert, dem Komplex aus DNA und Proteinen, aus dem die Chromosomen im Zellkern bestehen. Chromatin hat lebenswichtige Funktionen in einer Vielzahl von Prozessen, von denen die meisten durch ein komplexes System von kovalenten Modifikationen an Histon-Proteinen ablaufen.

Muster dieser Modifikationen sind wichtige Informationsträger, deren Weitergabe über die Zellteilung hinaus an beide Tochterzellen besonders wichtig für die Aufrechterhaltung des Zellzustandes im Allgemeinen und des Transkriptionsprogrammes im Speziellen ist. Die Entdeckung von epigenetischen Vererbungsphänomenen — mitotisch und/oder meiotisch vererbbare Veränderungen von Genfunktionen, hervorgerufen durch Veränderungen an Chromosomen, die nicht auf Modifikationen der DNA-Sequenz zurückzuführen sind — war bemerkenswert, weil es die Hypothese widerlegt hat, dass Informationen an Tochterzellen ausschließlich durch DNA übertragen werden.

Die Replikation der DNA erzeugt eine dramatische Störung des Chromatinzustandes, welche letztendlich ein partielles Löschen der gespeicherten Informationen zur Folge hat. Um den epigenetischen Zustand zu erhalten, muss die Zelle Teile der parentalen Muster der Histonmodifikationen durch Prozesse rekonstruieren, die noch immer sehr wenig verstanden sind. Eine plausible Hypothese postuliert, dass die verschiedenen Kombinationen der Lese- und Schreibdomänen innerhalb von Histon-modifizierenden Enzymen lokale Umschreibregeln implementieren, die letztendlich das parentale Modifikationsmuster der Histone neu errechnen. Dies geschieht auf Basis der partiellen Informationen, die in der Hälfte der vererbten Histone gespeichert sind.

Es wird zunehmend klarer, dass sowohl Informationsverarbeitung als auch computerähnliche Berechnungen omnipräsent und in vielen Bereichen der Naturwissenschaften von fundamentaler Bedeutung sind, insbesondere in der Zelle. Dies wird exemplarisch durch die zunehmend populärer werdenden Forschungsbereiche belegt, die sich auf computerähnliche Berechnungen mithilfe von DNA und Membranen konzentrieren. Jüngste Forschungen suggerieren, dass sich Chromatin während der Evolution in eine mächtige zelluläre Speichereinheit entwickelt hat und in der Lage ist, eine große Menge an Informationen zu speichern und zu prozessieren. Eukaryotisches Chromatin könnte also als ein zellulärer Computer agieren, der in der Lage ist, computerähnliche Berechnungen in einem biologischen Kontext auszuführen. Eine theoretische Studie hat kürzlich demonstriert, dass bereits relativ simple Modelle eines Chromatincomputers berechnungsuniversell und damit mächtiger als reine genregulatorische Netzwerke sind.

Im ersten Teil meiner Dissertation stelle ich ein tieferes Verständnis des Leistungsvermögens und der Beschränkungen des Chromatincomputers her, welche bisher größtenteils unerforscht waren. Ich analysiere ausgewählte Grundbestandteile des Chromatincomputers und vergleiche sie mit den Komponenten eines klassischen Computers, mit besonderem Fokus auf Speicher sowie logische und arithmetische Operationen. Ich argumentiere, dass Chromatin eine massiv parallele Architektur, eine Menge von Lese-Schreib-Regeln, die nicht-deterministisch auf Chromatin operieren, die Fähigkeit zur Selbstmodifikation, und allgemeine verblüffende Ähnlichkeiten mit amorphen Berechnungsmodellen besitzt. Ich schlage deswegen eine Zellularautomaten-ähnliche

eindimensionale Kette als Berechnungsparadigma vor, auf dem lokale Lese-Schreib-Regeln auf asynchrone Weise mit zeitabhängigen Wahrscheinlichkeiten ausgeführt werden. Seine Wirkungsweise ist demzufolge konzeptionell ähnlich zu den wohlbekannten Theorien von komplexen Systemen. Zudem hat der Chromatincomputer volatilen Speicher mit einem massiven Informationsgehalt, der von der Zelle benutzt werden kann. Ich schätze ab, dass die Speicherkapazität im Bereich von mehreren Hundert Megabytes von schreibbarer Information pro Zelle liegt, was ich zudem mit DNA und *cis*-regulatorischen Modulen vergleiche. Ich zeige weiterhin, dass ein Chromatincomputer nicht nur Berechnungen in einem biologischen Kontext ausführen kann, sondern auch in einem strikt informatischen Sinn. Zumindest theoretisch kann er deswegen für jede berechenbare Funktion benutzt werden. Chromatin ist demzufolge ein weiteres Beispiel für die steigende Anzahl von unkonventionellen Berechnungsmodellen.

Als Beispiel für eine biologische Herausforderung, die vom Chromatincomputer gelöst werden kann, formuliere ich die epigenetische Vererbung als rechnergestütztes Problem. Ich entwickle ein flexibles Simulationssystem zur Untersuchung der epigenetische Vererbung von individuellen Histonmodifikationen, welches auf der Neuberechnung der partiell verlorengegangenen Informationen der Histonmodifikationen beruht. Die Implementierung benutzt Gillespies stochastischen Simulationsalgorithmus, um die chemische Mastergleichung der zugrundeliegenden stochastischen Prozesse über die Zeit auf exakte Art und Weise zu modellieren. Der Algorithmus ist zudem effizient genug, um in einen evolutionären Algorithmus eingebettet zu werden. Diese Kombination erlaubt es ein System von Enzymen zu finden, dass einen bestimmten Chromatinstatus über mehrere Zellteilungen hinweg stabil vererben kann. Dabei habe ich festgestellt, dass es relativ einfach ist, ein solches System von Enzymen zu evolvieren, auch ohne explizite Einbindung von Randelementen zur Separierung differentiell modifizierter Chromatindomänen. Dennoch ängt der Erfolg dieser Aufgabe von mehreren bisher unbeachteten Faktoren ab, wie zum Beispiel der Länge der Domäne, dem bestimmten zu vererbenden Muster, der Zeit zwischen Replikationen sowie verschiedenen chemischen Parametern. Alle diese Faktoren beeinflussen die Anhäufung von Fehlern als Folge von Zellteilungen.

Chromatin-regulatorische Prozesse und epigenetische Vererbungsmechanismen stellen ein komplexes und sensitives System dar und jede Fehlregulation kann bedeutend zu verschiedenen Krankheiten, wie zum Beispiel der Alzheimerschen Krankheit, beitragen. In der Ätiologie der Alzheimerschen Krankheit wird die Bedeutung von epigenetischen und Chromatin-basierten Prozessen sowie nichtkodierenden RNAs zunehmend erkannt. Im zweiten Teil der Dissertation adressiere ich explizit und auf systematische Art und Weise die zwei Hypothesen, dass (i) ein fehlregulierter Chromatincomputer eine wichtige Rolle in der Alzheimerschen Krankheit spielt und (ii) die Alzheimersche Krankheit eine evolutionär junge Krankheit darstellt. Zusammenfassend finde ich Belege für beide Hypothesen, obwohl es für erstere schwierig ist, aufgrund der Komplexität der Krankheit Kausalitäten zu etablieren. Dennoch identifiziere ich zahlreiche differentiell exprimierte, Chromatin-assoziierte Bereiche, wie zum Beispiel Histone, Chromatin-modifizierende Enzyme oder deren integrale Bestandteile, nichtkodierende RNAs mit Führungsfunktionen für Chromatin-modifizierende Komplexe oder Proteine, die direkt oder indirekt epigenetische Stabilität durch veränderte Zellzyklus-Regulation beeinflussen.

Zur Identifikation von differentiell exprimierten Bereichen in der Alzheimerschen Krankheit benutze ich einen maßgeschneiderten Expressions-Microarray, der mit Hilfe einer neuartigen Bioinformatik--Pipeline erstellt wurde. Trotz des Aufkommens von weiter fortgeschrittenen Hochdurchsatzmethoden, wie zum Beispiel RNA-seq, haben Microarrays immer noch einige Vorteile und werden ein nützliches und akkurates Werkzeug für Expressionsstudien und Transkriptom-Profiling bleiben. Es ist jedoch nicht trivial eine geeignete Strategie für das Sondendesign von maßgeschneiderten Expressions-

Microarrays zu finden, weil alternatives Spleißen und Transkription von nicht-kodierenden Bereichen viel verbreiteter sind als ursprünglich angenommen. Um ein akkurates und vollständiges Bild der Expression von genomischen Bereichen in der Zeit nach dem ENCODE-Projekt zu bekommen, muss diese zusätzliche transkriptionelle Komplexität schon während des Designs eines Microarrays berücksichtigt werden und erfordert daher wohlüberlegte und oft ignorierte Strategien für das Sondendesign. Dies umfasst zum Beispiel eine adäquate Vorbereitung der Zielsequenzen und eine genaue Abschätzung der Sondenspezifität. Mit Hilfe der Pipeline wurden zwei maßgeschneiderte Expressions-Microarrays produziert, die beide eine umfangreiche Sammlung von nicht-kodierenden RNAs beinhalten. Zusätzlich wurde ein nutzerfreundlicher Webserver programmiert, der die entwickelte Pipeline für jeden öffentlich zur Verfügung stellt.

Acknowledgments

Writing the acknowledgments at the end of a long four-year journey full of ups and downs is a momentous occasion. Intellectually it has been the biggest challenge in my life so far and I put all the academic skills I learned during the last 10 years to this work. The following pages are significant because this thesis reflects those around me as much as it does myself. Dedicating so much time and energy to a project like a dissertation always bears the risk of losing sight to the non-academic world. I feel very fortunate that I was often surrounded by people that reminded me of the existence of the real world and the fun in it. Thank you to everyone who has given me insight, pleasure, and motivation and therefore helped to contribute to the completion of this thesis. In particular:

My supervisor Peter F. Stadler (I am tempted to say "Herr Professorin" but apparently this was all a big misunderstanding, wasn't it?) for your support, your understanding, and the opportunity to work on this thesis. I am still fascinated by your expertise, your incredibly fast way of thinking, and your often blisteringly fast email response time. I remember a particular instance when I wrote you an email at 3 a.m. on a Saturday night (sad, I know, but science never sleeps) and five minutes later you already answered me with an actual intellectually challenging response rather than a pure "Leave me alone, it is 3 a.m. at night!"— or was it just your doppelgänger that we all think must exist somewhere?

My second supervisor Sonja J. Prohaska for contributing tremendously to the success of this thesis. Although she faced a difficult time, I am deeply impressed with the way she managed to deal with this situation. I am particularly grateful for her incredibly creative mind in terms of how to visualize complex scientific processes and relationships. Thanks for your permission to modify some of your Figures that are included in this thesis!

Various people from the Fraunhofer Institute for Cell Therapy and Immunology, in particular Kristin Reiche for the wonderful and intense collaboration during the last two years or so, and Michael Specht for the development of the annotation database that was used in Chapter 6.

Charles L. Nunn, for shaping me during my time at the Max Planck Institute for Evolutionary Anthropology and my year in the US at Harvard University. I am very happy that we became friends, and I am tremendously grateful that you gave me the opportunity to work with you in Cambridge, and a lot of my academic skills originate from the time spent with you. I sincerely hope that we can meet again soon, no matter where in the world!

Petra for organizational support—filling forms in Germany never was and never will be something I want to do by choice unless I have to. Petra was always a tremendous help. At least for me, it seems that the knowledge of how to fill out a form is extremely volatile and is only stored in the short term memory. I also thank Jens for his technical support. He always helped me very quickly and his email response time often approached that of Peter's.

Various people in the lab for their friendship, especially Anne, Axel, Christian, Henry, Irma, Jens, Mario, Lydia, and Steffi (names are in alphabetical order to avoid any unintentional ranking). Axel, thank you for going bouldering with me so regularly! A special thanks goes to Irma and Steffi —

sharing an office with you for four years was really a pleasure, and I very much enjoyed it! I strongly believe that we were one of the hardest working offices. I remember days when we were all present at 9 pm in the evening, and two out of three were still present after midnight. The couch in our office also deserves a special thanks. Be it power napping, general paper reading often followed by falling asleep, or actual sleeping at night — you were always very comfortable!

The members of my former "Studententruppe" (Alex, Anne, Christoph, Stefan) with whom I had the great pleasure to stay friends with during the last eight years or so. Our regular Beyerhaus meetings were always a great distraction from work and general pleasure. A special thanks is devoted to Christoph for all the collective moaning and grumping and our permanent attempts to stay positive and optimistic despite being snowed under with work. Thanks for your friendship, and come back to Böhlen (do not confuse with Polen) soon!

Nicole, for always supporting me during the first four years of my academic path and for accepting my chaotic working times. A special thanks to listening to one of my talks and providing valuable feedback!

The members of the "Chaotentrio" — Philipp and Oliver — as well as Bartschi, Christian, and Tobias, all of who contributed significantly to the fact that I did not go berserk during the last third of my dissertation by doing some of the things that men like most — barbecue, beer, PlayStation, partying, and eating unhealthy junk food. I strongly believe that some of the more alcohol-intense evenings recreated capacity in my brain for additional dissertation-related neurons — externally-stimulated Darwinism in the brain, as I tend to call this phenomenon. I also want to apologize that my sincere time limitations made it sometimes very difficult to remain trained in *Fifa 12*.

Club Mate, chocolate and sweets in general, all of which was crucial for staying awake during the endless late-working nights during the last half year or so. I also want to thank my genetics and epigenetics (or to God if any creationist happens to read this) that I do not need too much sleep to remain functional and productive. Lastly, I want to acknowledge the series "The Big Bang Theory" because every time I watched it, I realized that I am not alone with the everyday academic peculiarities. The following quote from Sheldon Cooper is analogously true for epigenetics: "The physics is theoretical but the fun is real".

My girlfriend Stephanie who was always incredibly understanding of my workaholic and somewhat slightly chaotic life. She was extremely supportive, carefully watched my excessive *Club Mate* consumption and often acted as a very competent and fast offline version of `http://dict.leo.org`. I sometimes drove her to the edge of insanity but fortunately she did not fall. Thanks for cooking for me. It is mainly because of you that I kept relatively healthy eating habits. Almost every day, after the magic words "Ich gehe jetzt ins Bett'l" followed by the mandatory "noch zwei Minuten Rückenkrabbeln", my night shift began. I learned important skills during these endless nights such as working as silently as possible, sneaking out of the room without waking her up and stumbling upon various items located almost randomly in the room, making euphemistic statements about what time it is, and finding creative reasons why the ice cream suddenly vanished over night (it's better for the teeth to eat it all at once instead of gradually, don't you know that?).

The proofreaders Thomas Arendt, Seth Barribeau, Mitchell O'Brian, Alexander Georgiev, Sten Heinze, Collin McCabe, Charles L. Nunn, Stephanie Proft, Sonja Prohaska, Kristin Reiche, Kimberley and Stephen Reef, Martin Smith, and Peter Stadler, all of who provided invaluable feedback that significantly improved the overall quality as well as the general readability and professionalism of

this thesis. If it was financially, physically, and time-wise possible, I would like to drink one beer for each corrected error with all of you (beer is cheap in Germany, you know).

Last but not least and with particular emphasis my whole family who always supported me unconditionally. Although they do not really understand what I actually do, they always believed in my success and the fact that they are proud of me and my past accomplishments always kept me motivated. As they are more comfortable with the German language, I will temporarily switch to German so that they will at least fully understand a few sentences of this thesis.
Ich bedanke mich mit ganzem Herzen bei meiner ganzen Familie — meinen Eltern Yvonne und Sylvio, meinem Bruder Michael, meinen Großeltern Lunja und Volker sowie Annemarie und Jos, und meiner Uroma Marianne. Auch wenn ihr wohl noch immer nicht genau wisst, was ich eigentlich die letzten vier Jahre so gemacht habe im gefährlichen und großen Leipzig, habt ihr mich dennoch geduldig unterstützt und nie daran gezweifelt, dass ich irgendwann fertig werde. Ein ganz besonderer Dank geht dabei an meine Großeltern, die mir nicht nur einmal aus diversen Notsituationen, die das Leben so mit sich bringt, geholfen haben. Ihr wisst, was ich meine. Achja, liebe Mutti, auch ich bin sehr stolz auf dich und dein streberhaftes Abschneiden in den letzten Jahren, und würdige dies hier explizit.

Finally, for all actual readers of this thesis: Have fun reading, and do not hesitate to contact me if you have questions, critique, or general comments!

Christian Arnold, November 2013

> If you aren't in over your head, how do you know how tall you are?
> — *T.S. Eliot*

Contents

Abbreviations

AD	Alzheimer's disease
bp	base pair(s)
caRNA	chromatin-associated RNA
CC	chromatin computer
CEM	custom expression microarray
CRM	*cis*-regulatory module
DNA	desoxyribonucleic acid
EA	evolutionary algorithm
EST	expressed sequence tag
HAT	histone acetyltransferase
HDAC	histone deacetylase
HDM	histone demethylase
HMT	histone methyltransferase
IC	information content
kb	kilobase (1000 bp)
lncRNA	long non-coding RNA
ncRNA	non-coding RNA
PTM	post-translational modification
RNA	ribonucleic acid
TF	transcription factor
TM	Turing machine
UTR	untranslated region

List of Tables

List of Figures

Chapter 1

Motivation

1.1 The Increasing Significance of Chromatin and Epigenetics

Eukaryotic genomes are of unprecedented complexity, and we are just beginning to unravel the magnitude of the intricacy that underlies their genomic regulation. In the last years, our understanding of the significance of epigenetic mechanisms has steadily increased. Generally, epigenetics may be defined as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [1] although the term is often used for non-heritable processes, as exemplified by the definition "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" [2].

Eukaryotic genomes are organized as chromatin, the complex of DNA and proteins that forms chromosomes within the cell's nucleus. Its basic structure consists of proteins called histones with DNA wrapped around them, together referred to as nucleosome. Chromatin generally controls DNA accessibility and has pivotal roles in the cell for regulating gene expression, maintaining cell identity, genome packaging, and DNA damage repair. It also carries a partial "annotation" of genomic features. Indeed, it may be considered the "Swiss army knife" of biology, an all-in-one device suitable for every purpose. Collectively, the chromatin regulatory system displays high evolutionary variability and plasticity and consists of an incredibly complex network of diverse molecular players with dependencies among nearly all participants.

Eukaryotes have evolved a complex system of covalent chemical modifications of histones. These histone post-translational modifications (PTMs) play a particularly important role in a plethora of biological tasks. For example, H3K9me3 (i.e., trimethylation of the histone protein H3 at position 9, which is a lysine residue) is highly correlated with transcriptional repression and stably silences genes.

1.2 Chromatin as a Biological Computer

Computers are typically immediately associated with the traditional silicon-based computer technology that is omnipresent in our daily lives. However, more and more information processing and computation are discovered as fundamental processes of many fields [3]. Natural computing in particular is a fast-emerging, diverse, and fascinating area because it originates from naturally occurring biological systems. Ever since their initial discovery, natural computing techniques inspired the development of novel problem-solving techniques, for example by evolutionary optimization and swarm intelligence algorithms. Intriguingly, it becomes increasingly clear that computing may not only be regarded as an artificial science but also as a natural one [3–5]. An excellent example for computational processes that are observed in nature is molecular computing, which tries to both use molecules for computation and understand the information processing and computational nature of molecular processes in general.

Cells, the "building blocks of life", are incredibly complex and highly sophisticated biological units with huge information processing capabilities. Researchers therefore regarded the cell or specific cellular components repeatedly and with increasing frequency as a cellular computer capable of performing complex biological (*in vivo*) "computations" [4, 6–14]. Indeed, advanced information processing with molecules inside living cells is omnipresent and occurs on all scales. It can be found, for example, in complex structures such as the brain [15], in regulatory and signaling pathways within cells, and even within single biomolecules [16].

Work from Prohaska et al. [12] suggests that during evolution, chromatin has been converted into a potentially powerful computational device capable of storing and processing large amounts of information, caused by a number of key molecular inventions that substantially expanded the cell's regulatory scope [12]. Consequently, it has recently been postulated repeatedly that eukaryotic chromatin may act as a computational device capable of performing "computations" in a biological context [12, 13, 17]. Indeed, a recent theoretical study showed that a simple model of chromatin computation, very similar to that proposed in Prohaska et al. [12], is computationally universal and hence conceptually more powerful than gene regulatory networks, for example [13].

1.3 Epigenetic Inheritance

A set of phenomena termed epigenetic inheritance is particularly important for the maintenance of cell identity. Its discovery was a remarkable and unconventional finding because it disproved that information is passed to daughter cells only through DNA. Epigenetic inheritance phenomena can principally be divided into (i) somatic or mitotic epigenetic inheritance and (ii) transgenerational or meiotic epigenetic inheritance. The former thus concerns information transfer across cell divisions and the latter across organisms. For both types, a diverse set of putative theories have been

proposed for how such mechanisms work mechanistically. Because the existence and significance of transgenerational epigenetic inheritance in higher eukaryotes in particular is still highly debated and questioned [18], I will hereafter only focus on somatic epigenetic inheritance.

To maintain cellular identity, epigenetic patterns must be stably inherited across cell divisions. However, DNA replication constitutes a dramatic disruption of the chromatin state that effectively amounts to partial erasure of stored information. To preserve its epigenetic state the cell reconstructs (at least part of) the histone PTMs by means of processes that are still very poorly understood. A plausible hypothesis for the propagation of patterns of histone PTMs across cell divisions is that the different combinations of reader and writer domains in histone-modifying enzymes implement local rewriting rules that are capable of "recomputing" the desired parental patterns of histone PTMs on the basis of the partial information contained in that half of the nucleosomes that predate replication.

This recomputation-based model is based on positive feedback loops in nucleosome modification. Its popularity is based on its simplicity and experimental support that for various histone PTMs, histone-modifying enzymes bind to modified histones of the same type with higher affinity [19–26]. The existence of such recruitment-based conversions have been demonstrated for a number of histone PTMs such as H3K9/H3K27 methylation and H4K16 deacetylation [27–29].

Reconstituting the parental patterns of histone PTMs by histone-modifying enzymes to maintain epigenetic stability may be regarded as a chemical reaction system. A property of all biological systems is their intrinsic stochastic nature, and explicitly including stochasticity in the underlying model greatly enhances physical and biological accuracy. Indeed, due to the exponential increase in computing power, stochastic *in silico* modeling of such chemical reaction systems has emerged as a realistic and powerful means to improve understanding of these highly complex systems.

The recruitment model has furthermore been particularly intensively analyzed, both analytically and computationally, with important insights into the dynamics of the system and their overall potential and limitations. However, all approaches are subject to various limitations that limit their explanatory power such as insufficient realism in the modeling of the corresponding chemical reaction system [30–35] and highly simplified mathematical descriptions of the system [36–39]. Additionally, analyzing the potential of stably inheriting more complicated patterns of histone PTMs, which are likely to occur in nature simply due to the sheer complexity of histone PTMs and their crosstalk (histone code), has yet to be performed.

1.4 The Significance of the Chromatin Computer in Alzheimer's Disease

Maintaining correct epigenetic patterns throughout the lifetime of an organism is crucial for cellular stability and identity [40]. Misregulation of epigenetic (inheritance) mechanisms often has fatal consequences that may contribute significantly to various diseases. For example, chromatin has been increasingly associated with cancer [41] and Alzheimer's disease (AD) [42, 43]. The latter is the most common, irreversible form of dementia with no effective treatment being available so far [44]. It is mostly diagnosed in elderly people of age 65 or older [45]. The number of AD cases is estimated to triple by 2050 to over 100 million worldwide due to population aging [46]. It also poses a major challenge for health care systems due to its long duration, its current incurability, and its degenerative and terminal nature, with estimated costs of over $600 billion per year [46].

AD is characterized by severe memory loss, an impairment of other cognitive functions, and a substantial overall loss of brain volume [47]. A large array of putative causes has been postulated, and a multitude of genetic studies indicate that AD is not caused by a simple mutation because only a very small percentage of AD cases can be linked to mutations in specific genes. Intriguingly, the role of epigenetics in the etiology of AD is increasingly being recognized [48–54] although the causality of epigenetic changes in AD have yet to be established [51].

AD seems to be evolutionarily young disease and is currently believed to occur only in humans. No other mammalian species recapitulates all of the key features of AD [55] although some are also susceptible to Alzheimer-like symptoms [55, 56]. Data availability for non-human primate species is very rare and it therefore remains unclear to what extent Alzheimer is indeed a human-specific disease. An intriguing hypothesis concerns whether the genomic loci that are differentially regulated in AD are similarly evolutionarily young (i.e., whether they show signs of recent changes in their genomic structure). Additionally, if chromatin plays decisive roles in AD, one would expect to find an abundant number of differentially expressed chromatin-associated transcripts.

To identify disease-associated loci, biological experiments involving genome-wide transcriptome profiling must be performed. One specific technology for that purpose are microarrays. They have been used ubiquitously in biomedical research for over a decade now for applications such as single nucleotide polymorphisms detection, chromatin immunoprecipitation, gene expression, and DNA methylation. They provide a way to quantify the expression of thousands of nucleic acid samples (targets) by hybridizing them to known sequences (probes) in a massively parallel and often genome-wide manner in a single experiment. Despite the emergence of more advanced high-throughput methods such as RNA-seq, microarrays still offer some advantages and will remain a useful and accurate tool for expression measurements [57]. Most microarray providers offer custom expression microarrays (CEMs) for which the represented RNA transcripts can be precisely defined. CEMs are increasingly popular because they are more cost-effective than tiling arrays and offer more

flexibility. Although microarray technology is relatively mature, the compilation and preprocessing of target sequences prior to probe design is non-trivial. Due to the pervasiveness of transcription in mammalian genomes [58–61], various design issues for CEMs are often neglected although they may have a profound impact on microarray data analysis and statistical validation. Thus, the development of methods for the design of high-quality CEMs is important to improve the reliability, accuracy, and interpretability of expression measurements.

1.5 Objectives and Outline

The overall objectives of this thesis are twofold. First, I aim to elaborate on the notion that chromatin may be regarded as a cellular computer able to perform biological computation. I also aim to formulate epigenetic inheritance as a computational problem. To the best of my knowledge, these computational aspects of chromatin have remained largely unexplored. Second, I aim to identify differentially expressed loci in AD, with a particular focus on chromatin-associated transcripts to address the hypothesis that a dysregulated chromatin computer plays important roles in AD. Additionally, I assess whether AD may be considered as an evolutionarily young disease. To the best of my knowledge, these two hypotheses have yet to be addressed explicitly in a systematic fashion.

This thesis will be laid out as follows. First, in Chapter 2, I introduce the chromatin-based regulatory system and epigenetic inheritance and give required biological background. I discuss the significance of epigenetics for gene regulation, with a particular focus on candidate players such as histone PTMs and proposed models for epigenetic inheritance. I also highlight the consequences of cell division for epigenetic inheritance and review existing computational and analytical models for epigenetic inheritance. The remainder of the thesis is split into two parts.

In the first part, the focus is on the notion that chromatin may be regarded as a cellular computer able to perform biological computation. In Chapter 3, I specifically elaborate on that notion, analyze its major components, and investigate similarities and differences to ordinary computers. I compile properties of the CC, its mode of action, estimated memory size, computational power, and other characteristics. This chapter is mainly based on:

> *Arnold C, Stadler PF, Prohaska SJ. 2013. The Eukaryotic Chromatin Computer: Tasks, Components, Properties, Computational Power. in preparation.*

In Chapter 4, I investigate if epigenetic inheritance can be considered as a computational problem and if so, whether the CC is organized in a way that is amenable to the solution of this problem. For this, I implement a flexible and chemically and biologically accurate stochastic simulation system for the study of recomputation-based epigenetic inheritance of individual histone PTMs. I also analyze which parameters are most decisive for the stability of epigenetic states across cell divisions, and assess to what extent the potential power of chromatin computation is harnessed in real biological

systems. The chapter is mainly based on the following publication:

> *Arnold C, Stadler PF, Prohaska SJ. 2013. Chromatin Computation: Epigenetic Inheritance as a Pattern Reconstruction Problem. Journal of Theoretical Biology 336(7): 61-74.*

In the second part of the thesis, I first discuss various previously neglected issues in the design of CEMs that may have a profound impact on microarray data analysis and statistical validation. Specifically, in Chapter 5, as a prerequisite for the chapter that follows, I describe a bioinformatics pipeline that has been developed for the design of high-quality CEMs in the post-ENCODE era. I also describe the `Alzheimer Custom Array`, one of two CEMs that have been produced with the help of this pipeline. The chapter is based on the following publication:

> *Arnold C, Externbrink F, Hackermüller J, Reiche K. 2013. Design of Custom Expression Microarrays in the Post-ENCODE Era. Bioinformatics. submitted.*

In Chapter 6, I identify differentially expressed loci in AD, based on the `Alzheimer Custom Array`. I particularly focus on chromatin-associated loci and explore the significance of the CC in AD. I also assess whether AD may be considered as an evolutionarily young disease and what role chromatin regulation may play. The chapter is mainly based on:

> *Arnold C, Stadler PF, Hackermüller J, Reiche K, Überham U, Arendt T. 2013. Widespread and Diverse Differential Expression of Chromatin-Associated Transcripts in Alzheimer's Disease. in preparation.*

Finally, in Chapter 7, I summarize and conclude the results of this thesis. The Appendices provide supplementary material that is helpful for the interested reader.

1.6 Author Contributions and Note on the Use of Personal Pronouns

Impersonal style used to be required in academic writing, but this convention has changed. The use of personal pronouns is now common and encouraged. In this thesis, I follow this style. I will therefore continue using the first person pronoun "I" even though some parts represent the collective work of multiple individuals as stated above in the publication entries (e.g., parts of Chapter 4, Chapter 5, and Chapter 6).

I hereby explicitly acknowledge all individuals who contributed significantly to parts of this thesis. I also want to state and emphasize that the usage of the first person pronoun "I" does in no way discredit the individual author contributions.

Chapter 2

Chromatin Regulatory Mechanisms and Epigenetic Inheritance

2.1 Epigenetics, Chromatin, Chromatin Regulatory Mechanisms, and Transcriptome Complexity

2.1.1 Defining Epigenetics and Epigenetic Phenomena

Historically, the meaning of the term epigenetics has shifted multiple times [62]. Conrad Waddington first coined it in 1942 by fusing the word "genetics" with "epigenesis" to generally describe the interactions between genes and their surroundings to produce a phenotype [63, 64]. He thereby grounded epigenetics in a developmental context, and inherent in the original meaning of the word was the "view that epigenetic mechanisms are reset (that is, erased and re-established) at one point in the lifecycle of an organism" [65, p. 396].

Until the mid-1980s / early-1990s, a lack of knowledge of specific epigenetic mechanisms for genetic activity meant every scientist used his own definition of the term [62]. Robin Holliday, for example, defined it as "the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms" [66]. In the mid-1990s, with the realization that inheritance may not only be DNA-based due to the discovery of inheritable DNA methylation patterns, researchers defined epigenetics as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [1], which is still used widely today.

With the realization that histone post-translational modifications (PTMs) (see Section 2.1.4.1) play crucial regulatory roles that correlate with transcription, various authors often used epigenetics for non-heritable phenomena, as exemplified by the definition "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" [2]. David Shechter[1] defines epigenetics as "information content that increases the complexity of the genome without changes in

[1]Assistant Professor in the Department of Biochemistry of the Albert Einstein College of Medicine, NY, USA

the gene sequence"[2]. Finally, in 2008, researchers agreed on a consensus definition of the term "epigenetic trait", defining it as a "stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" [67] but it remains to be seen whether this will generally be accepted. In summary, the term epigenetics is nowadays often employed loosely, inconsistently, and sometimes synonymously to "epigenetic inheritance" [68].

In this thesis the last definition will also be adopted although the term epigenetics is often used in the context of "maintaining stable states of gene expression" [67] without involving chromosomal changes. For example, in medicine, the proteosome and particularly transcription factor (TF) networks are often called "epigenetic". Undoubtedly, TFs alone may play pivotal roles for the inheritance of expression states, and it seems conceivable that the transcriptional status of chromatin, rather than the specific histone PTM patterns or other chromosomal changes, may be transmitted during cell division [69–71]. The maintenance of particular expression states then depends on sufficient quantities of the TF that may be established by positive feedback loops to maintain its own expression in the absence of the original signal that once activated it [69]. Indeed, transient transcriptional errors caused by a single altered transcript in a TF may cause heritable phenotypic changes by reprogramming the transcriptional network [72].

A detailed description of epigenetic phenomena is out of the scope of this thesis and covered in myriads of excellent reviews (e.g., [73]). However, they include, but are not limited to, position-effect variegation, X-chromosome inactivation, paramutation, genomic imprinting, and gene silencing.

2.1.2 Chromatin Regulatory Mechanisms

Eukaryotic genomes organize as chromatin in the cell's nucleus. With the notable exception of dinoflagellates [74], the basic structure of chromatin consists of proteins (called histones) with ≈147 bp of DNA wrapped around them in ≈1.6 superhelical turns, together referred to as nucleosome (Figure 2.1B). The nucleosome is the most fundamental repeating unit of chromatin, composed of an octamer of histones (two copies of H2A, H2B, H3 and H4, respectively). Each histone has two domains: a globular domain that forms the nucleosomal core around which the DNA wraps [75], and long disordered tails that protrude from the nucleosomal core (Figure 2.1B). Indeed, their intrinsic disorder is crucial for their various functions [76]. Notably, both domains are structurally and functionally distinct [75]. The linker histone H1 binds the nucleosome at the entry and exit sites of the DNA and locates outside of the nucleosome. In contrast to the other histone proteins, it has no histone fold. Emerging evidence suggests that histone H1 PTMs also have various important functions (e.g., H1.4K34ac is involved in transcriptional activation and H1 may also regulate DNA methylation and histone H3 methylation) [77–79]. More generally, H1 seems to have important functions in establishing and maintaining higher-order chromatin structures.

[2] `http://www.shechterlab.org/2009/science/chromatin-and-the-histone-code`, last accessed in June 2013

During both DNA replication (see Section 2.2) and transcription, histones temporarily loosen their association with DNA and therefore cause a disruption of the nucleosome structure. Notably, histones may be bound by various proteins that aid and regulate chromatin assembly and disassembly as well as histone import into the nucleus (denoted histone chaperones).

An array of relatively regularly spaced nucleosomes (âĂŸbeads-on-a-stringâĂŹ fiber with a diameter of 11 nm) that, however, can change their precise location (nucleosome positioning) densely cover the genome. Among cells, nucleosomes may differ in their spacing and occupancy [82]. Various factors dynamically regulate nucleosome positioning, which is at least in part DNA-dependent [83, 84]. However, nucleosomes do not cover all genomic regions — active promoters and particularly enhancers frequently have nucleosome depleted regions, and these areas are especially accessible for TFs and more generally chromatin remodelers (reviewed in [84]). Furthermore, various factors tightly regulate nucleosome stability [85]. For example, newly assembled histones frequently replace "old" histones (histone turnover or exchange), which is particularly relevant for H2A and H2B [85–88]. In *Drosophila* and possibly also in human, histone turnover rates are even higher for active regions, epigenetic regulatory elements, and replication origins [89]. Nucleosomes can also be degraded (nucleosome eviction), replaced by newly assembled nucleosomes, composed of various histone variants (see Section 2.1.4.5), all of which adds yet another layer of regulatory complexity.

For the genomic DNA to fit inside the nucleus, nucleosomes are further hierarchically compacted to form various higher-order structures (e.g., the 30-nm chromatin fiber, Figure 2.1) that, however, remain only poorly understood [80, 90, 91]. These states have important roles for gene regulation and other biological processes because accessibility hinders with further compaction.

Chromatin also carries a partial "annotation" of genomic features such as promoters and enhancers [92], the exon/intron structure [93, 94], and the current state of transcription [95, 96]. Traditionally, researchers divided chromatin into two classes: eu- and heterochromatin. Whereas euchromatin is easily accessible, rich in genes and often associated with active transcription, heterochromatin is the opposite. However, Filion et al. [81] recently refined this crude binary classification (at least for *Drosophila*), and they identified five principal chromatin types (Figure 2.1 C).

Chromatin generally controls DNA accessibility and contributes to the recruitment of TFs. Indeed, chromatin can be considered the "Swiss army knife" of biology due to its utmost importance for a multitude of functions. It consists of an incredibly complex network of molecular players with dependencies among nearly all participants, and various factors collectively coordinate, regulate and maintain it. Therefore, chromatin can be seen as an advanced signaling module [97] that is a critical responder to external cues such as stress [98].

Lastly, chromatin-based regulation and epigenetic mechanisms display high evolutionary variability and plasticity. Their specific mode of action may thus be fundamentally differ among species, and any conclusions drawn from one particular species should be treated with caution.

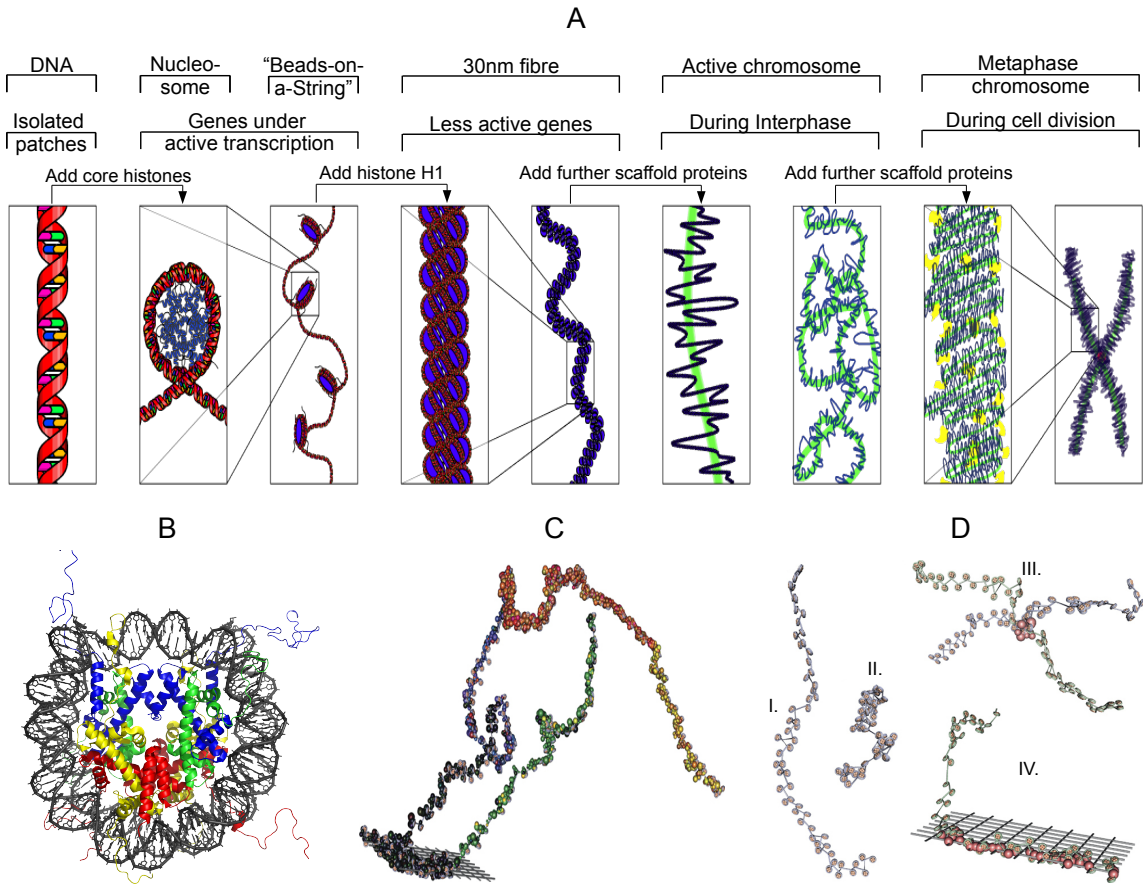


Figure 2.1: *Chromatin organization and higher-order structures. In A–D, fundamental features of the organization of chromatin and the various higher-order structures are shown. For details, see text. A and B are Wikimedia Commons files ("Chromatin Structures.png" and "Nucleosome 1KX5 colour coded.png", respectively), both of which are licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. C and D taken from Steensel [80].*
A: Overview of chromatin organization and higher-order chromatin structures. From left to right, the degree of compaction increases.
B: Crystal structure of an eukaryotic nucleosome core particle. For clarity, the coloring of individual histones and their protruding tails is distinctive, and the DNA wrapping around the nucleosome octamer is also visible.
C: Model of the five principal chromatin types in Drosophila *as identified by Filion et al. [81]. Two types that regulate different classes of genes correspond to the classical euchromatin, two to known types of heterochromatin, and the last type represents a previously unknown repressive chromatin type.*
D: Various principles of higher-order chromatin organization as illustrated by 3D computer simulations. `I` and `II` shows a nucleosome fiber consisting of 60 nucleosomes that is very flexible due to the variable degree of bending of the linker DNA and can adopt a range of configurations (`I`:extended configuration, `II`: more compacted configuration). `III` shows the frequent phenomenon of long-range chromatin interactions. Thus, spatially separated genomic regions come in close proximity due to the higher-order folding. `IV` illustrates another mechanism contributing to higher-order folding, namely interactions of particular genomic segments to fixed nuclear landmarks (gray lattice) such as the nuclear lamina, for example. These landmarks may act as anchoring sites for the chromatin structure.

2.1.3 Transcriptome Complexity and Pervasive Transcription

Eukaryotic genomes are incredibly complex, and we are just beginning to unravel some of the underlying mysteries. For example, since the completion of the human genome project, researchers now recognize that transcription is much more pervasive than previously imagined. The cell not only transcribes the 1.2% of the human genome coding for proteins but instead a total of ≈75% may be capable of transcription across cell lines [59], with a large number of overlapping transcripts (≈10–12 for each traditional protein-coding gene per cell line). Even more remarkably, 80% of the genome may be functional [99]. Furthermore, the gene regulatory network is highly complex and cell and tissue-specific with a large number of long-range interactions [100–103] and complex patterns of chromatin accessibility [103]. It is noteworthy, however, that these 2012 ENCODE Consortium publications have also generated harsh critique. Graur et al. [104], for example, argue that the authors use an extremely vague definition of function, isolate genomic analyses from their evolutionary context, and employ methods that consistently overestimate functionality.

The high complexity of the transcriptome is, however, indisputable and necessitates a reconsideration of what a gene is. Current attempts to define a gene therefore often categorize them by functional products instead of specific DNA loci. This is exemplified by the definition "a union of genomic sequences encoding a coherent set of potentially overlapping functional products" [105, p. 677].

For protein-coding and potentially also non-coding genes in eukaryotes, alternative splicing is a key mechanism to create diversity, and an estimated 95% of all multi-exon genes may be alternatively spliced [106]. The discovery of alternative splicing finally disproved the "one gene, one enzyme" hypothesis [107]. Splicing regulation is complex, and many factors control it [106, 108, 109]. Barash et al. [110] even proposed that a *splicing code* exists that describes how alternative splicing operates. Whereas alternative splicing is one of the most important mechanisms to generate mRNA structural complexity, its misregulation may lead to various diseases such as cancer [111, 112] and Alzheimer's disease (AD) [113, 114]. Splice sites can also be used as a conservation measure, as exemplified by a comparative analysis of the conservation of splice sites among mammalian species that demonstrated the evolutionary conservation of lncRNAs [115].

2.1.4 Selected Candidate Players in Epigenetics

2.1.4.1 Histone Post-Translational Modifications

Histone PTMs are covalent post-translational modifications deposited on histones by different histone-modifying enzymes. They are mostly present on the N-terminal tails protruding from the nucleosome (histone tail PTMs), which facilitates enzymatic access. The globular internal domain may, however, also be modified (histone core PTMs), and the majority of known histone PTMs located in the globular domain are lateral-surface histone PTMs (i.e., histone PTMs mapping to

Table 2.1: *Common nomenclature of histone PTMs. A common nomenclature names histone PTMs. It starts with the name of the modified canonical histone or histone variant, followed by the single-letter amino acid abbreviation and its position in the histone protein (relative to the N-terminal domain), an abbreviation of the modification type, and lastly the number of modifications and other modification details (such as symmetry information). The latter is currently only relevant for histone methylation because no other histone PTM is known to occur in multiple copies per residue. Particular residues may be subject to multiple histone PTMs although at most one distinct type of modification may occur at each residue.*

Histone PTM	Histone	Position	Modification type	Modification specifics
H2AXub	H2A.X	unspecified	ubiquitination (ub)	
H3K4me1	H3	Lysine at pos. 4	methylation (me)	monomethylation (me1)
H3K4me3	H3	Lysine at pos. 4	methylation (me)	dimethylation (me3)
H3K4ac	H3	Lysine at pos. 4	acetylation (ac)	
H3R2me2a	H3	Arginine at pos. 2	methylation (me)	asymmetric dimethylation (me2a)
H3R2me2s	H3	Arginine at pos. 2	methylation (me)	symmetric dimethylation (me2s)
H4K5cr	H4	Lysine at pos. 5	crotonylation (cr)	

residues in the globular domain that locate on the lateral surface of the histone octamer in close proximity to the DNA) [75].

Histone PTMs include various chemical alterations of lysine (acetylation, monomethylation, dimethylation, trimethylation, formylation, crotonylation [116], ubiquitination, sumoylation, biotinylation, succinylation [117], malonylation [117], propionylation [118], butyrylation [118], 5-hydroxylation [119], ADP ribosylation [120]), arginine (monomethylation, symmetric dimethylation, asymmetric dimethylation, deimination/citrullination), serine (phosphorylation, O-glycosylation/O-GlcNAcylation [121], hydroxylation), threonine (phosphorylation, O-glycosylation or O-GlcNAcylation [121]), tyrosine (hydroxylation [116]), glutamic acid (ADP ribosylation), and proline (isomerization) residues (Figure 3.5).

The discovery of at least 223 distinct[3] histone PTMs for the canonical histones H2A, H2B, H3, and H4 as well as H1.2 (191 if histone H1.2 PTMs are excluded, see Section 3.2.3 for more details) in humans alone revealed how fine-tuned and intricate the system of these covalent modifications is. To cope with this complexity, Turner [122] introduced the `Brno` nomenclature in 2005 [122] that is also consistently used throughout this thesis (Table 2.1).

Histone PTMs seem to be generally reversible because researchers already identified enzymes catalyzing the reversible reactions for almost all of them (e.g., decitrullination [123], desuccinylation [124], demalonylation [124], depropionylation [125]).

[3] counting the different forms of histone methylation separately because they may indeed have different functions

Researchers originally thought that histone PTMs always exist in a symmetrical fashion on both copies of the core histones in the nucleosomal octamer. However, nucleosomes can indeed be symmetrically and asymmetrically modified, and that these distinctions may even signal different biological outcomes [126, 127].

Histone methylation and acetylation are by far the best-studied modifications but studying other histone PTMs, such as histone phosphorylation, is increasingly common [128]. Confirmed by a wealth of experimental data, they play a decisive role in a wide array of biological processes such as regulation, cell differentiation, alternative splicing, DNA repair, stress response, development, cell cycle and mitosis, DNA replication, and various diseases [129–132]. Additionally, they can generally control the packaging of chromatin and therefore also DNA accessibility by altering either the positive charge of the histone tails (e.g., histone acetylation) or inter-nucleosomal interactions (e.g., histone methylation) [19, 116]. They can also either attract or inhibit chromatin binding complexes, causing subsequent regulatory changes [133]. Particular histone PTMs may also have functions outside of a nucleosomal context, for example by altering the binding affinities of histone chaperones (e.g., H4S47ph) [75]. Lastly, the linker histone H1 can also be covalently modified by various PTMs [116] but their functions are still elusive. Collectively, histone PTMs constitute a fine-tuned mechanism for regulating the structure and dynamics of chromatin. Their functional relevance is further supported by the observation that histone PTMs show signs of evolutionary conservation [134].

Generally, the functions of histone PTMs seem to be location-dependent. Histone tail PTMs mostly act as signaling factors that have no or only limited direct impact on the structure of chromatin, and specialized binding proteins carry out the function [75]. Histone core PTMs and particularly lateral-surface histone PTMs, however, seem to have a more direct structural and functional effect [75, 116, 135] because they alter intranucleosomal histone-DNA interactions. This may be achieved, for example, by forcing local increases in DNA unwrapping or by altering the affinity of the DNA-ïż£histone octamer. Consequently, they can influence nucleosome stability and mobility, accessibility of nucleosomal DNA to regulatory factors (e.g., H3K56ac). Lateral-surface histone PTMs may thus have a functional or even causative role for various regulatory events, such as transcription (e.g., H3K122ac), even in the absence of specific binding proteins, contrary to histone tail PTMs that may only be a non-causal byproduct of transcription. For example, a point mutation in H3K27 may cause a failure of transcriptional repression of genes normally repressed by *Polycomb* repressive complex 2 (PCR2) [136]. However, other histone PTMs, such as H3K4me3 (a hallmark of actively transcribed regions), may not necessarily be functionally important because transcriptional regulation may occur even in its complete absence [137].

The cell already establishes various histone PTMs on the H4 tail during synthesis of new histones. Such pre-existing histone PTMs seem to be required for chromatin assembly and DNA damage response signaling [138], and maybe even for epigenetic inheritance of particular histone PTMs

[139]. Their establishment after histone segregation may depend, for example, on sequence-specific factors through non-coding RNAs (ncRNAs) or TFs [140–142].

Importantly, most histone PTM do not function independently of one another. Instead, they act in concert to establish highly specific cellular signals. That is, the combination of different histone PTMs specifies unique downstream functions that can be interpreted by the cell. Researchers now commonly refer to this as the histone code hypothesis [143]. The histone code has the potential of massive complexity; however, only a small subset of all possible combinations seems to be used by the cell. For example, in human, Wang et al. [144] detected a common backbone of 17 distinct histone PTMs that colocalize and associate with promoters and enhancers. Since the description of the histone code hypothesis, researchers repeatedly observed crosstalk between two histone PTMs within and among histones of the same [145–148] and different nucleosomes in proximity (reviewed in [149, 150]). Examples include H2Bub and H3K4me2/H3K4me3 at actively transcribed genes [151, 152], H3S10ph and H4K16ac to mediate transcription elongation [153]. Furthermore, H3K9me and H3S10ph [154, 155] as well as H3S10ph and H3K14ac [156, 157] are linked.
Crosstalk may also be specific to particular cellular processes. For example, Latham et al. [158] observed mitosis-specific crosstalk for histone and non-histone proteins (specifically, between H2BK123ub and methylation of *Dam1*, a kinetochore protein). The finding that histones can be symmetrically and asymmetrically modified further expands the histone code [126, 127].
Due to the sheer number of histone PTMs and effector modules, crosstalk among more than two histone PTMs seems likely but current knowledge is still coarse, fragmented and incomplete due to the limited capacity to directly measure combinations of histone PTMs [145]. However, recently developed methodologies to systematically analyze combinatorial histone PTMs may shed new light on the magnitude of histone PTM crosstalk [159, 160].

Because histone PTMs mostly act and function in combination, correlating a single histone PTM with a definite functional outcome is difficult [161]. However, some broad generalizations can be made. For example, H4K20me3 marks centric (constitutive) heterochromatin, H4K36ac is important for transcription elongation, H3K27me3 associates with stable gene silencing mediated by the *Polycomb* complex, H2AS139ph has important functions in DNA repair [162], and H3K9me3 is a mark of transcriptional repression (reviewed, for example, in [141, 163, 164]). Furthermore, H3K36me has important roles in transcription elongation [132], alternative splicing [165], DNA repair [166], cell cycle and mitosis [132]. Due its diverse functions, it seems unsurprising that dysregulation of H3K36me levels associates with various diseases [132]. Another vital histone PTM is H4K20me, with functions in DNA repair (e.g., H4K20me2 as marker for double-strand breaks [167]), cell cycle regulation, DNA replication and chromatin compaction (reviewed in [168]). Lastly, H3K79me also plays a role in the cell cycle and mitosis in particular [169] but may also act as a marker or molecular timer for histone age [170]. Thus, histone PTMs often have multiple functions that may additionally be species-specific or even tissue-specific. For example, in addition to its function as marker for transcription start sites of active genes, H3K4me3 also has a protection

function against DNA methylation [141].

Histone PTMs play important roles for cellular memory and the re-establishment of regulatory programs after cell division [171]. The potential for a particular PTM to act as a heritable signal, however, depends on multiple factors:

- **The modified histone**
  H2A and H2B are much more mobile and more frequently exchanged than H3 and H4 [86, 172]. Thus, histone PTMs on H3 and H4 are more attractive candidates for epigenetic memory.
- **Modification type**
  Histone PTMs have strikingly different lifetimes and their deposition occurs at different rates. Acetylation events are measured in the order of minutes, whereas methylation events are stable for days [87, 173]. Thus, histone methylations are more stable than histone acetylations and therefore also more likely candidates for epigenetic inheritance, particularly because they predominantly occur on the H3 or H4 histone (see above).
- **Modification function**
  Francis [174] hypothesized that inheritance of histone PTMs is more likely for silent states than for active ones. Histone PTMs associated with active states, such as histone acetylations and phosphorylation, are often only transient signals that are set in response to a particular environmental stimulus (e.g., DNA damage caused by UV light).
- **Position on the histone**
  Particular histone PTMs located at the N-terminal tails (e.g., on histone H3) may be subject to position-specific phenomena such as histone tail clipping (i.e., the loss of amino acids) [175–177], which reduces the lifetime of histone PTMs and therefore its epigenetic inheritance potential. However, we know little about the frequency and significance of this process and its specific function. Nevertheless, histone clipping may have important roles for gene regulation and therefore also contributes to their inheritance potential. Furthermore, some residues, particularly in the globular domains of the histones in the core of the nucleosomes, may only sometimes be accessible (e.g., during transcription after nucleosome disassembly).

Only primary modifications, such as H3K9me3, may truly be epigenetic because they seem to be able to be inherited independently of the initial signal that triggered their formation, therefore contributing to cellular memory [178]. Secondary modifications such as most histone acetylations and histone sumoylation, however, require the initial trigger for their continuous presence or depend on either primary histone PTMs or other factors. They constitute dynamic signals to various cellular response pathways such as heat shock [179] and DNA damage. Indeed, the functional significance of histone PTMs in the processes they associate with is presently still unclear [75]. In particular, researchers still hotly debate whether histone PTMs and the presence of histone variants are a

cause or consequence of the transcriptional status [82, 178, 180, 181].

2.1.4.2 Higher-Order Chromatin Organization and Genome Topology

Inherent properties of mammalian genome topology are the existence of chromosome-specific territories and the arrangement of chromatin into local, megabase-sized chromatin interaction domains. They are highly conserved across species and stable across cell types [182, 183], with pluripotent stem cells having particularly distinct higher-order structures for increased robustness [184]. These topological domains do not seem to be a consequence of heterochromatin formation because they and particularly their boundaries appear to mark the end points of heterochromatin spreading [182]. Thus, they seem to act as crucial boundary elements to constrain H3K9me3 spreading.

Higher-order chromatin organization is likely to also play significant roles for epigenetic inheritance although we know little if and how higher-order structures are transmitted during cell division or whether particular higher-order structures are a cause or consequence of the transcriptional status and/or histone PTMs or other chromatin-related marks [90, 185, 186]. Higher-order chromatin organization is, however, associated with distinct histone PTM patterns [186] in pericentric heterochromatin and may promote or at least facilitate the spread of gene silencing [185]. Probst et al. [88] found that chromocentres (higher-order structures consisting of multiple pericentric heterochromatin domains) in mice are subject to only minor chromosomal rearrangements. They are mediated and regulated by *Polycomb* group proteins and may have crucial roles for the inheritance of repressive chromatin domains. Recent work found that cohesin complexes have important architectural roles for the establishment of higher-order structures and chromosome territories and therefore potentially also for epigenetic inheritance [187].

2.1.4.3 Non-coding RNAs

ncRNAs constitute a substantial portion of the transcriptome [59, 188]. They form a very heterogeneous, almost limitless versatile and abundant class of transcripts that can act both *in cis* and *in trans*. Their evolutionary conservation strongly indicates functionality [189]. Indeed, researchers associated them with a multitude of different functions [59, 190]. ncRNAs may achieve their intricate regulatory specificity by means of modularity, therefore assembling diverse protein combinations as well as interactions with the DNA and potentially also RNA [190].

Generally, ncRNAs can be divided in house-keeping or structural ncRNAs (e.g., tRNAs, snoRNAs, snRNAs, rRNAs) and regulatory ncRNAs. Due to their sheer diversity, the latter may be further subdivided into short ($<$ 50 bp such as miRNAs, siRNAs, and piRNAs), medium-long (50–200 bp such as promoter-associated RNAs), and long ncRNAs (lncRNAs) with a length of $>$ 200 bp

[191] although this length classification is somewhat arbitrary. lncRNAs show the greatest variety, with even further subdivisions into intergenic, intronic, UTR-associated, antisense, pseudogene, or enhancer-like ncRNAs. Generally, ncRNAs seem to have an almost boundless versatility. For example, ncRNAs may also exist as circular RNAs that are more stable and therefore less easily subject to degradation. They seem to form a large class of post-transcriptional regulators, and researchers already detected thousands of well-expressed, stable, and tissue-specific circRNAs [192]. circRNAs can, for example, counteract the function of miRNAs and therefore desuppress mRNA target expression [192, 193]. mRNAs, on the other hand, may also function as both mRNA and lncRNA [194]. Due to their functional diversity, cell type or tissue specificity, and frequent disease-association (see Chapter 6), they also offer great potential for the application as biomarkers, and researchers already use some of them [195, 196].

Epigenetic regulatory and inheritance roles of lncRNAs were first identified in the context of genomic imprinting (parent-specific gene expression) for the phenomenon of X-chromosome inactivation [197, 198]. lncRNAs are crucial for the silencing of the inactive X-chromosome (Xi) by forming repressive heterochromatin that prevents expression of most of the genes from Xi (reviewed in [199]). Prominent lncRNAs implicated in epigenetic inheritance include *HOTAIR*, *Kcnq1ot1*, *Airn*, and *Xist*, for example.

Numerous recent findings indicate that ncRNAs have important roles in modulating chromatin and its structure and therefore also for epigenetic inheritance (e.g., see [199–204]). Indeed, various lncRNAs can be bound by chromatin-modifying enzyme complexes such as *Polycomb* repressive complex 2 (PCR2). For example, Khalil et al. [201] showed that a large number ($\approx$20%) of lncRNAs play a role in the establishment of heterochromatin by binding *PCR2* (e.g., *HOTAIR*). Mondal et al. [188] came to a similar conclusion. They identified a number of intronic and intergenic regions in human fibroblast cells that harbor chromatin-associated RNAs. These regions additionally showed significant conservation across 44 mammals, thus strongly indicating functional significance.

These chromatin-associated RNAs may therefore have an important function in regulating gene expression by guiding chromatin-modifying complexes to particular genomic loci (Figure 2.2), for example those that must maintain an epigenetic memory [40, 185, 205].

lncRNAs typically act *in cis* but *in trans* regulation also seems to be common [199]. Lai et al. [206] showed that they may activate neighboring genes *in cis* by interacting with the co-activator complex *Mediator*. Interestingly, such enhancer-like lncRNAs seem to be a significant stimulator of the *Mediator* kinase activity towards H3S10ph, a histone PTM strongly associated with transcriptional activation [207]. lncRNAs may also prevent spreading of repressive histone PTMs [208].

In lower eukaryotes, another non-coding RNA phenomenon, namely RNA interference (RNAi), plays a role in epigenetic inheritance and even transgenerational epigenetic inheritance [209]. In yeast, for example, RNAi and small interfering RNAs (siRNAs) are crucial for the inheritance of

heterochromatin (i.e., H3K9me) [90, 210–214]. Specifically, RNA polymerase II shortly transcribes heterochromatin in the S phase during the cell cycle, and the RNAi machinery processes the resulting transcripts into siRNAs. In conjunction with other silencing factors, these siRNAs then recruit histone H3K9 methyltransferases to establish H3K9me. Additionally, DNA polymerase components help to mediate recruitment of the various epigenetic factors that are required for the faithful establishment of heterochromatin formation. They thus orchestrate DNA replication, RNAi, and histone methylation, which would also explain the cell cycle-regulated, RNAi-dependent heterochromatin silencing [213]. Cernilogar et al. [215] showed that in *Drosophila*, *Dicer* 2 and *Argonaute* 2, two key players in RNAi, globally associate with transcriptionally active loci by interacting with the core transcription machinery, thereby controlling the processivity of RNA polymerase II. However, the significance of RNAi mechanisms in higher eukaryotes remains unclear.

A legitimate question is why lncRNAs often function as epigenetic regulators, and not proteins or small RNAs. Indeed, lncRNAs seem to offer certain advantages as compared to proteins for epigenetic regulation [199]. First, they have the ability of allelic marking as a consequence of their tethering capabilities and their rapid turnover, in contrast to proteins (as their origin of transcription is lost when mRNA is shuttled to the cytoplasm) and small RNAs. Second, due to their large size, lncRNAs can specify a unique genomic loci (Figure 2.2), whereas TFs, despite their effective recruiting of regulatory factors, typically affect a large number of genes at once due to their recognition of relatively short and therefore abundant DNA motifs. Thus, lncRNAs and TFs in combination with chromatin-modifying enzymes may account for the required specificity (both in terms of space and time) during development [199].

2.1.4.4 Chromatin-Modifying Enzymes

Chromatin-modifying enzymes generally describe classes of macromolecules that physically associate with chromatin to modify and/or regulate chromatin structure, composition, and function. They may generally be divided into ATP-independent and ATP-dependent chromatin-modifying enzymes. ATP-independent complexes denote several distinct classes of enzymes able to modify histone proteins, whereas ATP-dependent complexes remodel chromatin to induce structural chromatin changes (e.g., reposition or deplete nucleosomes, exchange histones) by using the energy provided by the hydrolysis of ATP. Both have crucial roles for the maintenance of epigenetic states and epigenetic inheritance. *Polycomb* and *Trithorax* group proteins are particularly well-studied chromatin-modifying enzymes for the silencing and activation of gene expression, respectively, with pivotal functions as regulators of numerous developmental genes and more generally chromatin-remodeling [216]. Chromatin-modifying enzymes may specifically add/write (histone writer) or remove (histone eraser) histone PTMs (Figure 2.3 A). In eukaryotes, histone writers tend to be more numerous than the corresponding histone erasers [12]. Widely studied classes of chromatin-modifying, or

[4]Hereafter, the latter is used repeatedly throughout this thesis without repeated credit.

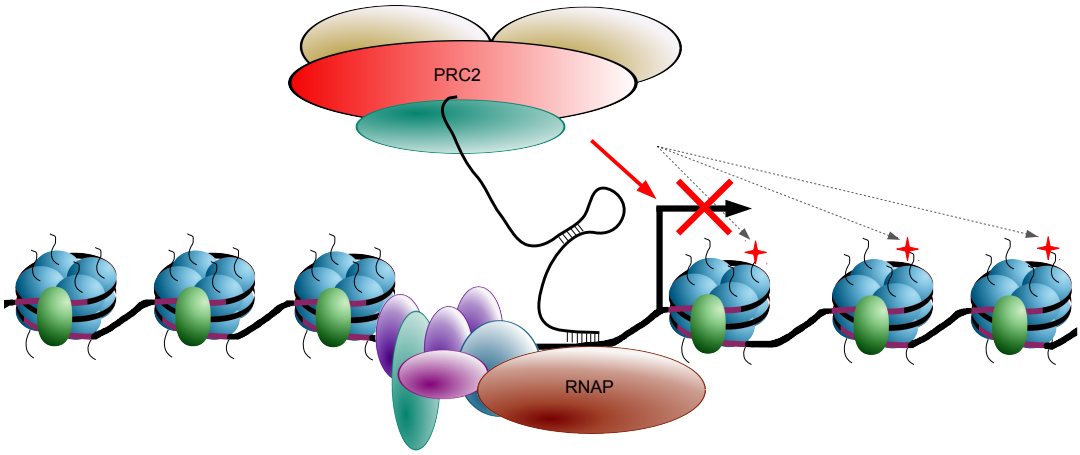


Figure 2.2: *Example of epigenetic regulation by lncRNAs. The figure shows a genomic locus that harbors a* cis*-acting lncRNA and the transcription elongation machinery. Upon transcription via RNA polymerase (RNAP), it can silence a surrounding gene by targeting and physically associating with* PCR2*. This subsequently methylates histone at position H3K27 co-transcriptionally (red stars at the protruding histone tails) and thereby silences the gene. For details, see text. The image is a composite image of parts of an image released to the public domain (transcription elongation machinery) and the Wikimedia Commons file "Nucleosome organization.png", which is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license*[4]*.*

more specifically histone-modifying, enzymes include histone acetyltransferases (HATs), histone deacetylases (HDACs), histone methyltransferases (HMTs), and histone demethylases (HDMs). HATs, for example, acetylate lysine residues on various histone proteins by transferring an acetyl group from acetyl-CoA to form ϵ -N-acetyl lysine, whereas HDACs[5] can perform the opposite reaction. Similarly, histone methyltransferases (HMTs) catalyze the transfer of one to three methyl groups from the cofactor S-Adenosyl methionine to lysine and arginine residues of histone proteins. Scientists initially believed that histone methylations are permanent (irreversible) but the discovery of a H3K4 demethylase showed that they can also be actively removed. One such example are *Jumonji* domain-containing proteins [217] such as *KIAA1718*, which can demethylate mono-, di-, and even trimethylated lysines.

Additionally, chromatin-modifying enzymes often associate with reader domains able to specifically recognize the presence or absence of particular histone PTMs (Figure 2.3). Such histone readers contain binding domains with a size of ≈50–150 residues that contain binding pockets for the recognition of individual histone PTMs (e.g., bromodomains for various acetylated H3 and H4 lysine residues, *PHD* domains for H3K4 methylation, *2014-3-3* domains for H3S10ph, see [218] for further examples) [145] (Figure 2.3 B).

[5]also more generally called lysine deacetylases (KDAC) because they also modify non-histone proteins

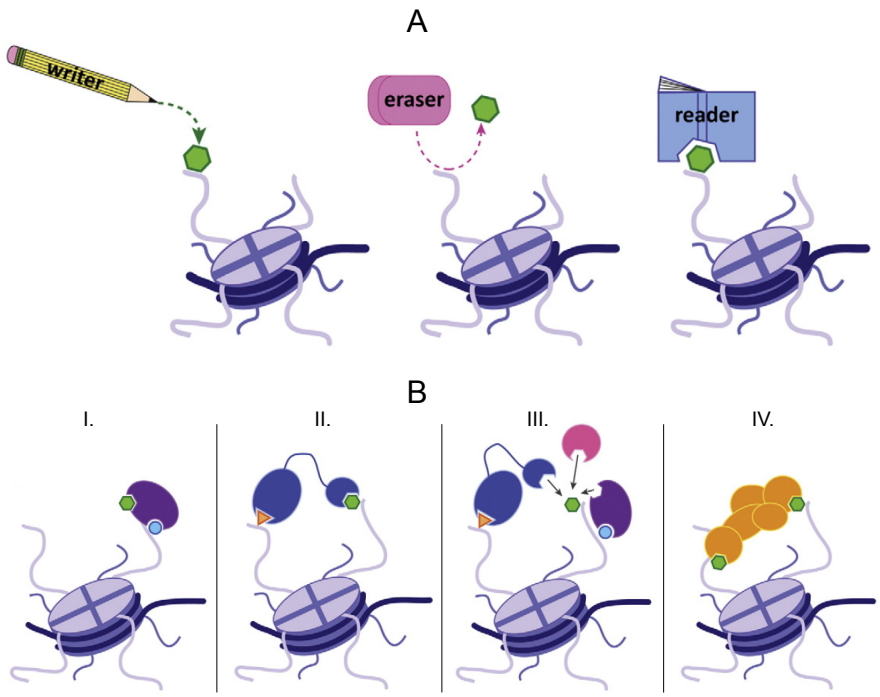


Figure 2.3: *Histone-modifying enzymes and their mode of action. Figures taken from Gardner et al. [219]. A: Schematic illustration of histone writers (left), erasers (middle) and readers (right), all of which can specifically or unspecifically associate with histone PTMs. See text for details.*
B: Mechanisms and diversity of histone reader modules. Binding of particular histone PTMs or their unmodified residues may occur in cis *on the same histone tail (I),* in trans *across histone tails (II), or even across nucleosomes (not depicted). Single histone PTMs may serve as a docking site for multiple proteins, with additional histone PTMs dictating the specific recruitment process (III). Lastly, histone-modifying enzymes often consist of multiple individual proteins that collectively form a multimeric complex and functionally distinct domains (IV, see also Figure 2.5).*

In contrast to TFs, many chromatin-modifying enzymes do not bind to specific DNA motifs and often even lack DNA-binding domains [200]. Consequently, they often bind more ubiquitously. However, binding specificity may be achieved through interactions with *cis*-acting ncRNA (see Section 2.1.4.3). Notably, numerous chromatin-modifying enzymes associate with the DNA replication machinery and their mode of action may therefore be intimately coupled to DNA replication (see Section 2.2). Similarly, these enzymes often associate with the transcription machinery [130, 181].

Histone writers can have very different mode of actions. HMTs, for example, can either be processive or non-processive (Figure 2.4). Most HMTs are processive (e.g., SET) but Frederiks et al. [220], for example, also discovered the non-processive HMT Dot1. Importantly, the kinetic mechanisms for processive and non-processive enzymes are very different. A non-processive mode of action, for example, implies that the different methylation states are not independent from one another, which may introduce functional redundancy (e.g., H3K79 methylation by Dot1) [170, 220]. Methylation

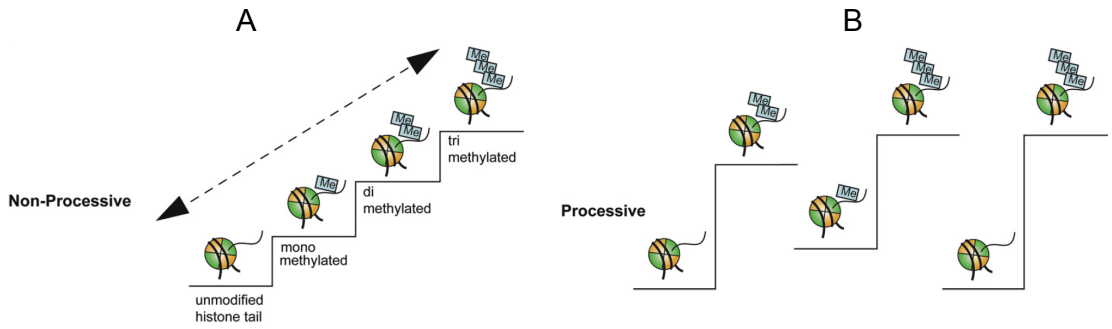


Figure 2.4: *Processive and non-processive histone-modifying enzymes. Methyl groups can be added in a step-wise and gradual fashion (i.e., catalyzing only one reaction at a time before releasing its substrate, non-processive, A) or in a continuous fashion by adding multiple methyl groups at a time and skipping one or two intermediate steps (processive, B) [222]. Figure taken from Scharf et al. [222].*

loss, on the other hand, is typically processive [221].

Histone-modifying enzymes are often part of large multisubunit complexes with almost arbitrarily complex combinations of reader, writer and eraser domains (Figure 2.5) [223]. Paired modules are particularly common for short-range histone crosstalk and include reader-reader, reader-writer and reader-eraser pairs [145]. Many multi-component chromatin-modifying enzyme complexes also contain multiple reader domains [24]. The chromatin-remodeling complex *RSC*, for example, consists of three and the *Polybromo* protein of six bromodomains (summarized in [218]). Various authors proposed that specific interactions among different histone PTMs play crucial regulatory roles, and subsequently extensively investigated to what extent the cell may achieve a combinatorial recognition of histone PTMs and how it functions mechanistically [223]. For example, multiple copies of reader domains may prefer cumulative effects of particular types of histone PTMs (e.g., histone acetylation) because their presence may lead to cooperative effects resulting from each individual domain [218]. However, single reader domains may also selectively recognize multiple histone PTMs. One such example is a single bromodomain from a testis-specific member of the *BET* protein family that is responsible for selectively recognizing diacetylated histone H4 tails [224].

Importantly, individual histone PTM readers may also be influenced by the modification state of adjacent residues. Thus, for effective binding and execution of their designated reaction, histone-modifying enzyme complexes may require a particular combination of histone PTM states (i.e., the presence and/or absence of multiple histone PTMs). Because recent data indicates the existence of both symmetrically and asymmetrically modified populations of nucleosomes [126], histone-modifying enzymes may also specifically recognize and/or require a symmetric or asymmetric modification state for a particular histone PTM. Such crosstalk is ubiquitous (see above and [150] for a review) and examples include the dual recognition of H4K20me2 and H2AK15ub by the bivalent histone PTM reader 53BP1 [225] and H2Bub-dependent H3K4 methylation by Set1 [151, 152]. Generally, such crosstalks can either be of sequential or combinatorial nature. Sequential interactions occur

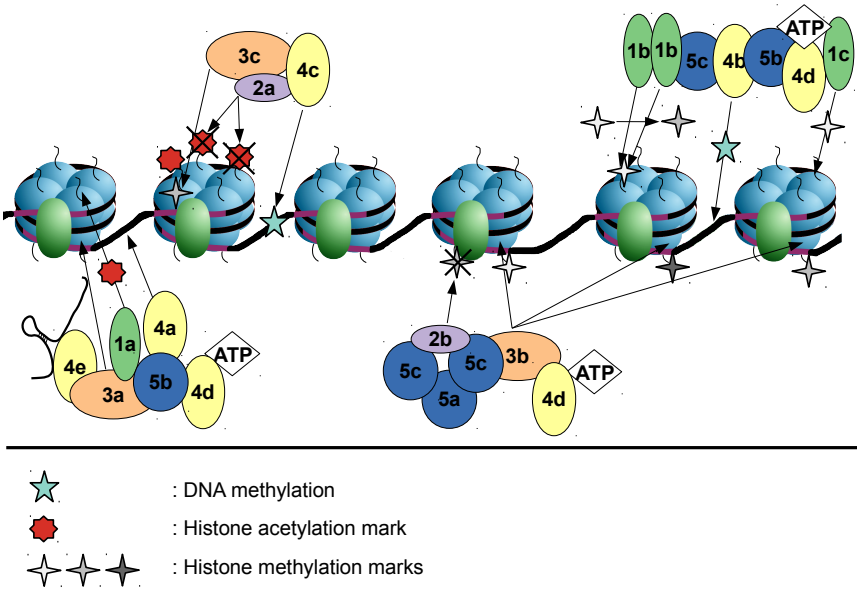
step-wise, as exemplified by a lysine methyltransferase that binds a particular histone tail through recognition of a histone PTM and subsequently deposits a second histone PTM. Combinatorial interactions, on the other hand, denote the recognition of multiple histone PTMs at once. Examples include antagonistic mechanisms where histone PTM readers recognize a particular PTM, which is, however, impaired if a second PTM is deposited [150]. Whether histone-modifying enzymes read individual histone PTMs simultaneously or sequentially, however, is still subject to intense research [145].

Chromatin-modifying enzymes often also link histone PTMs with other cellular processes or phenomena such as:

- DNA methylation (e.g., the HDM *LSD1* links H3K4 and DNA methylation [227])
- cell metabolism (e.g., the kinase *WEE1* links H2B phosphorylation and cell-cycle progression [228])
- environmental signals such as temperature (e.g., vernalization in flowering plants requires various HMTs, reviewed in [229])
- sex-determination (e.g., in mice, the H3K9 HDM *Jmjd1a* controls expression of the sex-determining gene *Sry* through the regulation of H3K9me2, and mice lacking *Jmjd1a* were subject to male-to-female sex reversal [230])

Various histone-modifying enzymes, such as kinases, methylases, and acetylases, depend on high energy co-substrates and can therefore be influenced in their level of activity by environmental and metabolite signals [161]. Intriguingly, histone-modifying enzymes from pathogens can also manipulate histone PTM patterns to promote their own survival by depositing novel histone PTMs. For example, a methyltransferase in the bacterium *Legionella pneumophila* may establish H3K14me3 to repress host gene expression and enhance its own intracellular replication [231].

Lastly, mutations in genes that code for chromatin-modifying enzymes may also cause severe diseases. For example, Zaidi et al. [232] found that $\approx$10% of severe cases of congenital heart disease may be caused by *de novo* mutations of heart-expressed chromatin-modifying genes with important developmental functions. Another prominent example is cancer, and researchers identified multiple mutations for various chromatin-modifying enzymes, such as somatic mutations of *UTX* (encoding a H3K27 demethylase), mutations of *EZH2* (encoding a H3K27 methyltransferase), and a translocation of *MLL* (encoding a protein that is recruited to many promoters and mediates H3K4 methyltransferase activity), that occur with high frequencies in multiple hematological malignancies (summarized in [233]). Therefore, understanding the role and importance of chromatin-modifying enzymes for the dynamics of chromatin states may also be important to improve understanding of "epigenetic" diseases (see also Chapter 6).



Symbol	Class	Functional description
1	HISTONE PTM WRITER	1a) Histone acetyltransferase activity
		1b) General protein lysine methyltransferase activity
		1c) H3K79 methyltransferase activity
2	HISTONE PTM ERASER	2a) Histone deacetylase activity
		2b) Histone demethylase activity
3	HISTONE (PTM) READER	3a) General H3 binding
		3b) Methylated H3 binding
		3c) H4K20me binding
4	BINDING DOMAINS	4a) DNA-binding
		4b) DNA-binding, methylase activity
		4c) Methylated DNA-binding
		4d) ATP-binding
		4e) ncRNA binding
5	CHROMATIN REMODELING AND SCAFFOLDING PROTEINS	5a) Chromatin remodeler
		5b) Helicase activity
		5c) Scaffolding protein

Figure 2.5: *Functional and structural complexity and diversity of chromatin-modifying enzymes and chromatin remodelers. The figure shows four different enzyme complexes for a genomic region composed of five nucleosomes. Each enzyme complex consists of several proteins and domains with their respective functions (Table below, which is adapted from Pu et al. [226]). As indicated, enzymes may either specifically or unspecifically recognize and/or modify histone PTMs and DNA methylation. Some complexes may even be important for higher-order chromatin organization and structural stability.*

2.1.4.5 Histone Variants

During evolution, various universal and lineage-specific variants of canonical histones have arisen. They acquired a diverse set of specialized functions in chromosome segregation, DNA repair, transcriptional regulation, sperm packaging, centromere maintenance [234], and sex chromosome

condensation [235, 236], for example. In contrast to their canonical paralogues, they are constitutively and not only replication-dependently synthesized. They mostly differ in only a few amino acids in their primary sequence [235] but can cause major structural differences, stability changes, and the addition or deletion of particular functional domains or histone PTMs, all of which generally may alter nucleosome dynamics.

In mammals, common variants of histone H2A include H2A.X, and H2A.Z, for example. The latter associates with the promoters of actively transcribed genes. Histones H4 and H2B so far have no known ubiquitously expressed variants, and researchers so far only identified a few tissue-specific variants of H2B (e.g., sperm-specific and testis-specific variants) [236]. A particularly important variant of H3 is CENP-A, which is a key determinant of centromere function and both necessary and sufficient for the epigenetic maintenance of centromers [237, 238]. Interestingly, even minor differences between histone variants may be specifically recognized (e.g., human H3.1 and H3.2 differ only at a single residue but nevertheless display differences in the abundance of particular PTMs) [239]. Lastly, particular histone variants, such as H3.3, also influence nucleosome disassembly dynamics because their incorporation may alter the likelihood of a split of H3-H4 tetramers (see Section 2.2.2).

Importantly, researchers repeatedly suggested that various histone variants (e.g., CENP-A, H3.3, H2A.Z, and H2A.X) are crucial for epigenetic inheritance (reviewed in [88, 236]) but except for CENP-A [234], the precise mechanisms remain elusive. For example, whether histone variants may be specifically recognized by particular enzymes or if their primary function is to induce structural changes of the surrounding chromatin is currently an open question [236]. Henikoff [85] hypothesized that the H3 variant H3.3 may be a carrier of epigenetic information, based on the finding that active transcription during interphase leads to higher levels of H3.3 incorporation, which after cell division then serves as signal for transcription. However, this model has only limited support and researchers could not yet validate its various assumptions. Fachinetti et al. [234] showed that CENP-A is indeed the epigenetic carrier of centromere identity that identifies, maintains and propagates centromere function through a conserved two-step mechanism. Thus, histone variants can be *bona fide* epigenetic marks.

2.1.4.6 DNA Methylation

DNA methylation generally denotes the conversion of the DNA base cytosine to 5-methylcytosine (5mC) by adding a methyl group. It is common in most eukaryotes with a large genome but can also be found in some small-genome eukaryotes and even bacteria. DNA methyltransferases (DNMTs), a set of very conserved proteins, mediate it. It is typically present within cytosine-guanine dinucleotide (CpG) islands [221, 240] around gene promoters and within gene bodies although it may also occur in a non-CpG context in, for example, embryonic stem cells [241]. It generally affects the transcription

of genes and therefore has crucial rules in gene regulation, carcinogenesis, allele-specific expression of imprinted genes, silencing of transposons, and X-chromosome inactivation, for example. Recently, however, an increasingly complex picture emerges between DNA methylation and gene expression (reviewed in [242]).

Various factors faithfully preserve DNA methylation during DNA replication and therefore constitutes a *bona fide* epigenetic mechanism. After DNA replication, DNA is originally only hemi-methylated because the newly synthesized daughter strand lacks any DNA methylation, whereas the other daughter strand carries the parental DNA methylation marks. In contrast to de-novo DNMTs, maintenance DNMTs can specifically recognize hemi-methylated DNA (e.g., *DNMT1*). After their recruitment to the corresponding genomic locations through *UHRF1*, they methylate DNA also on the other strand and therefore restore the parental modification pattern (reviewed in [178]). *DNMT1* preferentially associates with ubiquitylated histone H3, and *UHRF1* provides the mechanistic link between DNA methylation and DNA replication through H3 ubiquitylation [243].

Notably, DNA methylation interacts with almost all other putative candidate players of epigenetic inheritance, for example by recruiting chromatin-modifying enzymes [40, 242]. Due to its stable inheritance, it may act as primary signal for the (re-)establishment of histone PTMs (e.g., replication-coupled deposition of H3K9me2 with *DNMT1* and *UHRF1* [181]). However, several lines of evidence in genome-wide studies in cancer cells suggest that silencing of particular chromatin domains precedes DNA methylation but it remains unclear if this is generally true (reviewed in [242]).

Researchers originally thought that DNA methylation is irreversible but it is now clear that removal of DNA methylation is widespread and may occur via multiple active and passive mechanisms (reviewed in [221, 240]). This is particularly important for germline cells because genome-wide demethylation happens shortly after fertilization to reset the DNA methylation status.

Methylcytosines may be further converted to 5-hydroxymethylcytosine (5hmC), which can be catalyzed by the ten-eleven translocation 1 (*TET1*) enzyme, for example. Hydroxymethylation is likely to represent a distinct epigenetic state with important functions in transcriptional regulation during embryonic development [244], passive DNA demethylation [245], genomic imprinting, maintenance of cellular identity, epigenetic regulation of gene expression, and suppression of transposable elements [246, 247]. Researchers also reported further cytosine derivatives such as 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [247–249] although they are less abundant. Their functional and epigenetic significance is also presently unclear [250]. So far, researchers established their involvement in the cytosine demethylation pathway. For example, after oxidation of 5mC and 5hmC to 5caC, 5caC can be specifically recognized and excised [248]. They may also play a role in transcriptional regulation by reducing the rate and substrate specificity of RNA polymerase II transcription [250].

2.2 DNA Replication and Mitosis

2.2.1 DNA Replication

DNA replication is a complex but relatively well-understood process. Therefore, due to space limitations, only details relevant for the maintenance of epigenetic states will be presented.

In the synthesis phase of the cell cycle during mitosis, the cell replicates its DNA in a semi-conservative manner. DNA replication has an intrinsic strand bias because it only occurs in the 5' to 3' direction. Whereas the leading strand replicates in a continuous fashion, the lagging strand synthesizes discontinuously.

Notably, the genome replicates at different times. For example, transcription of active genomic loci occurs earlier than transcription of silent loci. Indeed, the specific replication timing program is evolutionary well conserved [251]. Replication timing in turn correlates with levels of chromatin compaction and therefore associates vaguely with the replicated chromatin type (e.g., eu- and heterochromatin) [178, 252–254]. This has led to the idea that different loci replicate in dependence of their transcriptional status, which may establish the decisive component for potential epigenetic inheritance mechanisms (reviewed in [178]).

During each DNA replication, chromatin must undergo dramatic perturbations due to the melting of the DNA double helix, and this wave of disruption poses major challenges for any chromatin-based inheritance mechanism. For example, histones must temporarily dissociate from the original position during passage of the replication fork. Histones (or histone oligomers) then distribute (segregate) between the two daughter strands and are deposited in new nucleosomes. Due to the crucial importance for epigenetic inheritance, models for nucleosome disassembly and histone segregation are discussed in more detail below.

2.2.2 Nucleosome Disassembly

During DNA replication or transcription, the nucleosome octamers disassemble (temporarily) to allow access to DNA, followed by immediate reassembly after DNA replication. Multiple models of nucleosome disassembly are theoretically possible but as depicted in Figure 2.6, only two of them have experimental support: (i) splitting of the nucleosome octamer into two individual H2A-H2B dimers and the H3-H4 tetramer (model 1) and (ii) further tetramer splitting of H3-H4, resulting in H2A-H2B and H3-H4 dimers (model 5). Thus, nucleosome disassembly occurs in a stepwise manner, starting with the removal of H2A-H2B dimers and followed by the H3-H4 tetramer [255]. Histones H3-H4 predominantly segregate as tetramers and only rarely as dimers [172, 181, 256]. The relative frequencies for these two models seem to depend mainly on whether the assembly machinery incorporates histone variants. For example, splitting of $(\text{H3.1-H4})_2$ is rare compared to

$(H3.3\text{-}H4)_2$ tetramers [172, 256]. Because H3.3 mainly enriches in euchromatin, an alternative formulation for the relative frequencies of the two models is whether active or passive regions replicates.

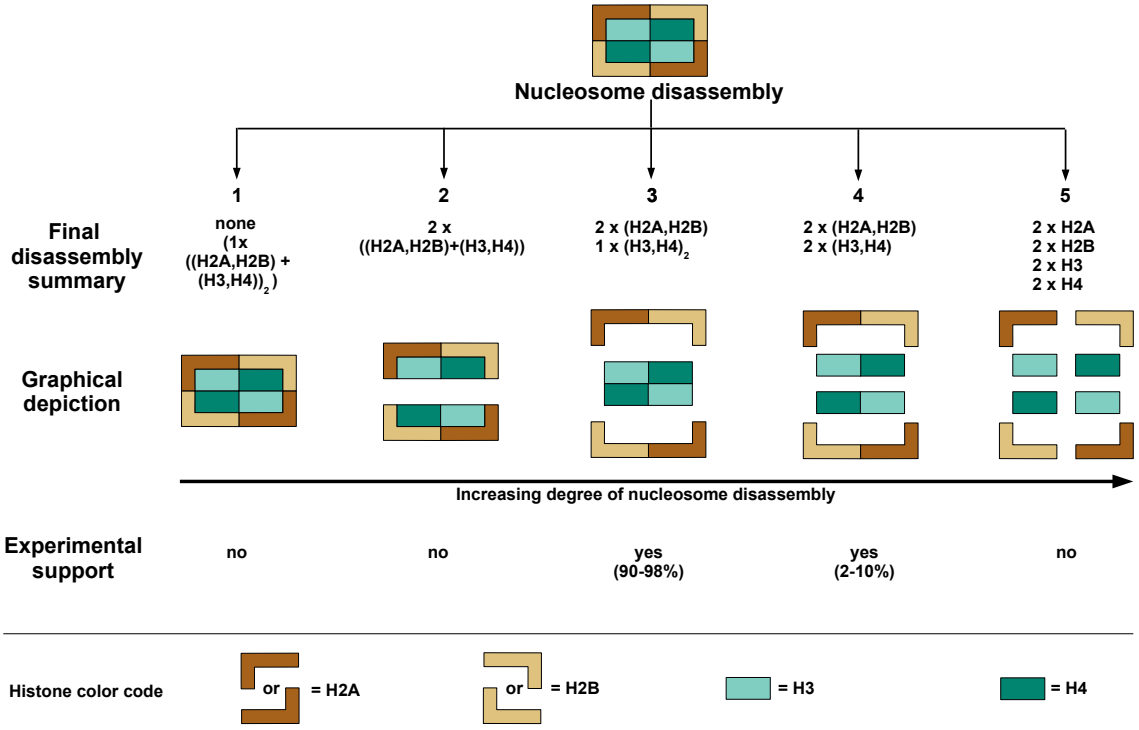


Figure 2.6: *Overview of different nucleosome disassembly models. Five different models that are theoretically possible are presented. For each model, a summary of the final disassembly, a graphical depiction, and its experimental support is provided. For clarity, the octameric nucleosome is color-coded (see bottom). The values for the relative frequencies of the two models with experimental support were taken from [172, 256]. For more details, see text.*

2.2.3 Mitosis

Mitosis is the first event in the M phase of the cell cycle and denotes the separation of previously replicated chromosomes into two identical sets of chromosomes. After mitosis, each chromosome set is in its own new nucleus, and cytokinesis[6] follows immediately. During mitosis, the nuclear envelope dissolves and chromatin and the nuclear architecture undergo massive and global reorganization. It also globally silences transcription and ejects sequence-specific TFs and therefore theoretically also provides a window of opportunity for remodeling gene expression and epigenetic patterns [71]. Thus, mitosis can be seen as a labile state susceptible to global gene expression changes.

[6] cytoplasmic division at the end of mitosis or meiosis

2.2.4 Histone Segregation

Histone segregation denotes the distribution of the individual histone oligomers to the two daughter strands following replication. From a systematic point of view, models of histone segregation may principally be classified according to two different criteria: (i) whether histones segregate as dimers or tetramers and (ii) whether histone segregation occurs in a random or non-random fashion (Figure 2.7). As indicated in Figure 2.7, researchers proposed various models in the literature, and the three models with experimental support are now described in more detail.

Random model
The random model assumes that parental histones segregate randomly between both strands [26, 257, 258]. For each of the two strands at each position, the probability of incorporation of parental histones is therefore 50%. Because of the disassembly of nucleosomes into individual histone dimers and tetramers and its octameric structure, a joint deposition of parental and newly assembled histones is therefore frequent for newly formed nucleosomes. This results in a replication-coupled dilution of parental histone PTMs during subsequent DNA replication events if the cell does not specifically modify newly assembled histones to reconstitute the parental histone PTM pattern. This model has the best experimental support among all models [181, 259, 260]. Furthermore, it is relatively irrelevant whether H3-H4 complexes segregate as dimers or tetramers because the overall enrichment for a particular histone PTM is of primary importance [185].

Semi-conservative model
The semi-conservative model assumes that parental histones distribute equally between both daughter strands. Although it provides a simple and elegant inheritance mechanism because intranucleosomal templated modification copying events may theoretically easily restore the premitotic histone PTM patterns, the model requires an identical modification state within each nucleosome (i.e., identical modifications of the two copies each of the four core histones). However, only a minority of H3-H4 complexes segregate as dimers (see above) and nucleosomes may indeed exist in symmetric and asymmetric populations [126, 127], thus arguing against its existence.

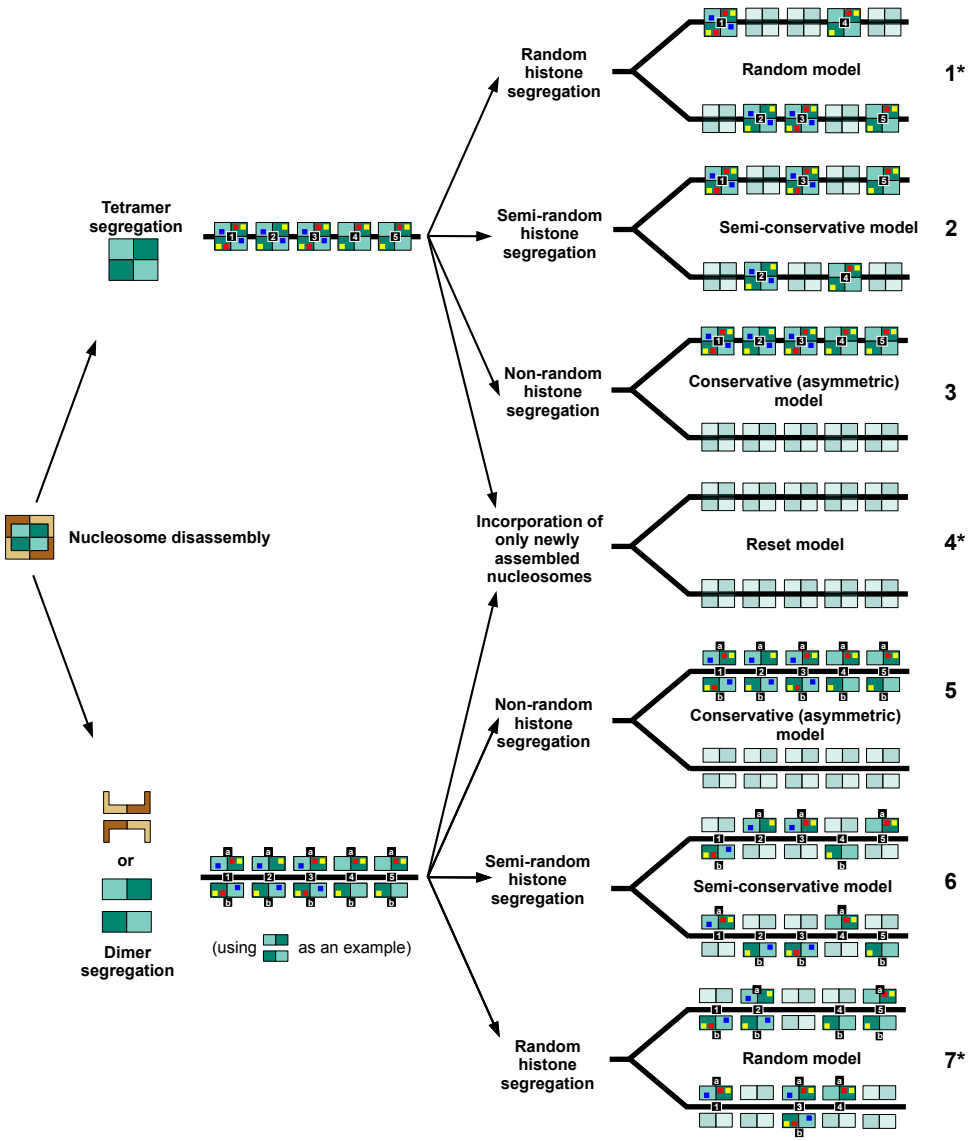


Figure 2.7: *Overview of different histone segregation models during DNA replication. A systematic overview of various possible histone segregation models is shown, based on a genomic region consisting of five nucleosomes (numbered from 1 to 5) and various histone PTMs (as indicated by the colored rectangles within the histones). Note that nucleosomes may be modified symmetrically (1–3) or asymmetrically (4–5). The models are divided into tetramer (1–2) and dimer (3–5) segregation models, respectively. Furthermore, models with experimental support are highlighted with a star next to the exemplary model number (1, 4, 7) and newly assembled histones or nucleosomes are drawn with transparency to distinguish them from their parental counterparts. For simplicity, (i) parental nucleosomes retain at their original position following DNA replication, (ii) only the histone oligomers relevant for the various models are shown and not the full nucleosomes, (iii) newly assembled histones are generally unmodified, (iv) only H3-H4 dimers for the dimer segregation models are shown although the model also applies to H2A-H2B dimers (as illustrated left), and (v) common names are also provided (e.g., see [88, 185]). To the best of my knowledge, however, nobody made the explicit distinction between dimer and tetramer segregation so far. The color coding of the histones is identical to Figure 2.6. For more model details, see text.*

Conservative or asymmetric model

The conservative or asymmetric model [88] assumes that parental histones distribute asymmetrically onto daughter strands. For example, one of the daughter strands may contain all parental and the other strand only newly assembled histones. Such a model may be possible by the intrinsic strand bias (see above). However, the re-establishment of premitotic histone PTM patterns requires not yet observed interstrand crosstalk. Such a model may nevertheless exist for specialized chromatin domains or particular time intervals but it seems unlikely that it plays a dominant role.

Reset model

Petruk et al. [261] proposed a model in which the cell only incorporates newly assembled histones (Figure 2.7). Despite the complete loss of information, inheritance of the patterns of histone PTMs may still be possible because the corresponding histone-modifying enzymes may remain bound during replication or quickly rebind afterwards (see Section 2.3.2.1 for more details).

2.2.5 Retainment of Parental Histones After DNA Replication

Whether parental histones reassemble close to their original loci from which they were evicted by the replication fork during DNA replication also has major implications for the stability of epigenetic information and epigenetic inheritance. Radman-Livaja et al. [260] addressed this important issue, which is mechanistically distinct from histone segregation. Using a quantitative model partly based on experimental data, they estimate that during replication, two-thirds of all histones reincorporate within ≈400 bp (which vaguely corresponds to two nucleosomes) of their prereplication locus. These values are in agreement with related previous observations and therefore suggest that although histones stay in close vicinity of their prereplication locus, segregation mechanisms are inherently stochastic and rather imprecise.

Radman-Livaja et al. [260] also followed the inheritance of parental histone H3 through multiple generations and surprisingly observed a tendency for parental histones to cluster preferentially downstream of the 5' ends of most genes (peaking around the +3 nucleosome, with respect to the promoter). Interestingly, the effect was strongest for long and poorly transcribed genes. With the help of a quantitative model, they find that the parental distribution of H3 histones fits well to the model if (i) histones reincorporate in proximity of their original locus during replication (see above), (ii) replication-independent replacement occurs and (iii) transcription-associated lateral movement or histone "passback" (histone displacement from 3' to 5') occurs. The last phenomenon contributes most to the net accumulation of parental histones near the 5' end of genes, and has an estimated value of ≈90 bp per cell cycle. However, as mentioned explicitly by Radman-Livaja et al. [260], because the majority of nucleosomes in yeast are well positioned [262], the transcription-dependent passback value cannot be taken literally and must be interpreted probabilistically. That is, "an octamer will be passed back in a given cell cycle in each cell — a passback value of ≈80 bp suggests

that there is a 50% chance that histones on a given gene will be shifted back one position towards the 5' end in a single cell cycle." [260, p.14].

However, the significance of these results remains to be determined. For example, experiments were done using yeast, and whether the mechanisms are also applicable to higher eukaryotes is presently unclear. The authors also note explicitly that all results must be interpreted with caution because the effect of "wild-type, untagged nucleosomes on the behavior of the epitope-tagged histones" [260, p.11] is unknown.

2.3 Epigenetic Inheritance

2.3.1 Definition, Differentiation, Evolutionary Implications

I already introduced epigenetic inheritance in Chapter 1, and in what follows, I want to provide an overview of some of the proposed models for epigenetic inheritance. I note again that the focus of this work is on somatic and not transgenerational epigenetic inheritance. Although the presented candidate players and models may also be applicable, transgenerational epigenetic inheritance encompasses additional key players and a highly specialized set of mechanisms not discussed here [18] such as PIWI-interacting RNAs [263] and other RNA molecules [264]. In addition to these two types of vertical transmission of epigenetic states, Molnar et al. [265] described RNA-mediated horizontal transmission in plants, which, however, will not be further discussed.

The realization of the existence of epigenetic inheritance phenomena and the possibility of environmentally-directed inheritance revitalized Lamarck's theory of inheritance of acquired characteristics (theory of adaptation). Subsequently, a wealth of studies discussed and requestioned the precise contributions of Darwinian and Lamarckian modalities of evolution and whether we have to alter our thinking about evolutionary change [65, 68, 266]. Various proposed models of evolution integrate epigenetic inheritance phenomena into general theories of evolution but much controversy remains with respect to the precise role and importance of epigenetic inheritance phenomena in shaping phenotypic variation and the course of evolution [267].

2.3.2 Models

Biological reality is extremely versatile and complex, and epigenetic inheritance cannot be explained by any single mechanism. Instead, it must be seen as a set of diverse and non-mutually exclusive mechanisms and interconnected molecular pathways acting in combination to collectively restore the parental modification state. Indeed, virtually all models for epigenetic inheritance utilize multiple components of the chromatin regulatory system but the individual significance of the various epigenetic key players and their precise mode of action remains to be determined [69, 242].

In recent years, researchers proposed numerous divergent models for epigenetic inheritance [181, 268]. Generally, they all rely on particular chromatin-associated marks or signals to re-establish the premitotic chromatin state. Such marks can be very diverse and include, for example, the presence or even absence of the presented key players (e.g., histone variants, histone PTMs, DNA methylation, bound ncRNAs or proteins, chromatin-modifying enzymes).

For DNA methylation, a *bona fide* epigenetic inheritance mechanism exists (see Section 2.1.4.6). Similarly, for particular chromatin domains, specialized inheritance mechanisms exist that involve epigenetic key players. For example, the cell faithfully inherits centromeres epigenetically, as illustrated in Section 2.1.4.5. The inheritance of silent ribosomal DNA (rDNA, the DNA that codes for ribosomal RNA) may also occur via epigenetic mechanisms involving ncRNAs, histone PTMs, DNA methylation and chromatin-modifying enzymes. Specifically, the *NoRC* complex is crucial for the stable inheritance of silent rDNA by establishing heterochromatin structures at the silent rDNA region by recruiting DNA methyltransferases and HDACs to the rDNA promoter (reviewed in [269]). *PARP1*, a histone-modifying enzyme with NAD^+-dependent ribosyltransferase activity that can also establish ADP-ribosylation on histones, plays a particularly central role. *PARP1* and its enzymatic activity is critical for the maintenance of silent rDNA chromatin by associating with a subunit of the *NoRC* complex via the ncRNA *pRNA*, which ultimately leads to heterochromatin formation.

Inheritance of DNA methylation and particular specialized chromatin domains are replication-bound and represent faithful "copying" mechanisms. Inheritance of histone PTMs, on the other hand, is more diverse. Whereas templated copying mechanisms exist for some histone PTMs that are governed and facilitated by their own presence, other histone PTMs are re-established *de novo* after replication by various factors using histone-independent mechanisms [181]. The former therefore represent primary histone PTMs that may truly be epigenetic. However, even for primary histone PTMs such as H3K9me3, inheritance mechanisms are much more imprecise. As shown in Section 2.2.4 and Section 2.2.5, after DNA replication, no faithful *bona fide* mechanism seems to exist that re-establishes the premitotic histone PTM pattern on both daughter strands. Instead, because of the replication-coupled dilution of histone PTMs, the parental modification state has to be "recomputed" using more error-prone, replication-independent maturation mechanisms. For example, higher forms of H4K20 and H3K79 methylation are gradually re-established throughout the cell cycle [270, 271].

Therefore, the inheritance of histone PTMs does not occur in a pinpoint fashion, and the minimal unit of epigenetic inheritance is therefore likely to span multiple adjacent nucleosomes. The cell may therefore be able to recognize and interpret average modification levels of chromatin domains spanning several nucleosomes. For example, De Vos et al. [170] suggested this for multiple histone PTMs such as H3K9me and H3K79me. Consistent with this, histone PTMs with putative epigenetic inheritance mechanisms associate with long domains (e.g., H3K9me3). One possible explanation for the existence of such imprecise mechanisms for histone PTMs is that the cell has to remain

somewhat flexible after DNA replication to provide a window of opportunity for remodeling gene expression patterns and to allow changes in the cell fate [71].

The presented mechanism for the inheritance of rDNA is also an excellent example of the potential significance of the specific replication timing (see Section 2.2) because contrary to silent rDNA that replicates in late S phase during the cell cycle, active rDNA replicates in early S phase [269, 272]. It also demonstrates how different actively transcribed and transcriptionally inactive chromatin may be inherited. Indeed, epigenetic inheritance mechanisms seem to differ fundamentally between active and repressive chromatin domains. For each of these two generic chromatin types, I now present one specific model with good experimental support.

2.3.2.1 Mitotic Retention of Proteins and RNA

A cellular âĂŸmemoryâĂŹ of the transcriptional program across cell divisions may be sustained by gene or mitotic bookmarking [71, 273–275], which denotes a collection of mechanisms that either facilitate the retention of proteins or RNA at previously active gene loci [275] or their fast post-mitotic transcriptional reactivation [274]. The former may therefore be called epigenetic according to the definition I use in this thesis due to the continuous binding of proteins or RNA. Generally, gene bookmarking involves a large variety of genetic and epigenetic players such as histone PTMs, histone variants, nucleosome architecture, TFs, sequence-specific DNA binding proteins, DNA topology, and ncRNAs.

One specific mechanism is the marking of active loci by promoter-associated H4K5ac and subsequent post-mitotic recognition by a protein called *BRD4*, which remains associated with the region through strong binding with H4K5ac [274]. After exit from mitosis and recruiting additional *BRD4* molecules, various factors gradually decompact chromatin in the vicinity of the promoters so that it becomes transcriptionally active. As noted by Wang et al. [171], histone PTMs may therefore "act as 'countermarks', 'landmarks', and 'bookmarks' to displace, recruit, and 'remember' the location of regulatory proteins during and shortly after mitosis" [171, p. 175].

Similarly, based on findings in *Drosophila*, Petruk et al. [261] proposed a model in which the cell only incorporates newly assembled histones but chromatin-modifying enzymes remain bound during replication or quickly rebind to restore the parental modification pattern, the former of which qualifies as epigenetic mechanism. Similarly, Blobel et al. [276] observed the retention of *MLL* (a histone methyltransferase), which accelerates transcription reactivation following mitotic exit. However, it remains unclear if histone PTMs alone constitute bookmarks, or if associated chromatin-modifying enzymes must remain bound or at least remain in vicinity [71].

2.3.2.2 Recruitment-Copying Model

One of the models for epigenetic inheritance with the best experimental (e.g., [29, 277]) and theoretical (e.g., [30, 34, 36, 38]) support is based on positive feedback loops in nucleosome modification [278, 279] (Figure 2.10). It is also known as mark-copying model [178], recruitment-copying model, signal reinforcement and spreading [64], and templated modification copying model [181]. Similarly to mitotic bookmarking, it also uses various epigenetic key players. According to the model, the premitotic histone PTM state may be gradually restored after DNA replication using parental histones with existing histone PTMs in the vicinity of the newly synthesized histones as a template [178] (Figure 2.8). The recruitment-copying model therefore belongs to the class of replication-independent maturation (recomputation) although the name may imply otherwise.

The model is popular because of its simplicity and the experimental support that for various histone PTMs, histone-binding enzymes bind to modified histones of the same type with higher affinity [19–26]. Such recruitment-based conversions also exist for H3K9 and H3K27 methylation [26] (the latter in the context of plant vernalization[7] in *Arabidopsis*, which is a classic epigenetic process that involves PRC2-based silencing of the floral repressor *FLC*, for example [280]) and H4K16 deacetylation [27–29].

A prerequisite for any recruitment-based inheritance model is the approximate reassociation of parental histones in the vicinity of their original locus from which they were evicted, which indeed seems to be the case, at least in yeast [260]. Hathaway et al. [277] recently determined the propagation rate for H3K9me3. Emerging from nucleation centers that carry H3K9me3, the authors showed that in *in vivo* mouse embryonic stem cells, H3K9me3 then propagates symmetrically (i.e., it spreads in both directions) and continuously at average rates of $\approx$0.18 nucleosomes per hour to produce domains of up to 10 kb.

Despite direct chemical and theoretical support, various authors question it, mainly because of the non-sufficient robustness of epigenetic states and the high nucleosome dynamics. However, as noted by Dodd et al. [38], the idea that such dynamics prevent histone PTMs being able to transmit epigenetic information misses the critical distinction between a system and its components because although individual components (i.e., nucleosomes) may be subject to multiple modifications, the system (i.e., histone PTM domains) can nevertheless achieve high stability (e.g., [36, 38] but see also Chapter 4).

The stability of histone PTMs in very short domains is difficult to conceive with the recruitment-copying model alone [178, 260] because they may easily get diluted throughout DNA replication (e.g., H3K4me3 at transcriptional start sites of most genes) [185]. Gene bookmarking mechanisms that integrate various components of the chromatin regulatory system therefore seem to be more appropriate for this task.

[7] ability to flower in the spring by the perception and memory of winter

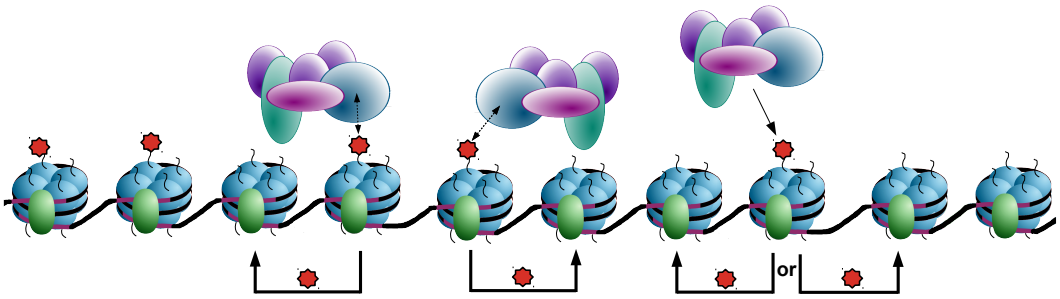


Figure 2.8: *Recruitment-copying model for the inheritance of epigenetic states. The main principle of the recruitment-copying model is illustrated, namely the templated recruited modification of newly synthesized nucleosomes based on neighboring, premitotic nucleosomes carrying histone PTMs. Positive feedback loops arise by histone-modifying enzymes that preferentially bind to modified nucleosomes of the same type (i.e., in the same state). Thus, nucleosomes with a particular histone PTM directly or indirectly recruit histone-modifying enzymes to establish the same histone PTM on nucleosomes in close vicinity [36]. For more details, see text.*

2.3.3 Findings and Predictions of Analytical and Computational Models

Due to their relevance for the remainder of this thesis, in the last section of this chapter, I summarize findings and predictions of analytical and computational epigenetic inheritance models. Notably, I will introduce some phenomena and properties often found in biological systems that are also important for epigenetic inheritance and models of epigenetic inheritance. To improve understanding of the various model predictions and results, I also provide a brief explanation of these concepts in Figure 2.10.

Using analytical or computational methods, researchers analyzed various models for epigenetic inheritance in recent years (reviewed in [17]). Although most of these analyses were purely theoretical, they provided important insights into the dynamics of the system and their overall potential and limitations. The recruitment-copying model attracted particular attention, both analytically [30–35] and computationally [36, 38, 39]. These studies all confirmed the potential of histone PTMs for stable epigenetic inheritance. All computational models were based on an influential paper by Dodd et al. [36] published in 2007, which provided several important and unexpected insights. The authors used a very simple mathematical model based on positive feedback loops in nucleosome modification (see Section 2.3.2.2), loosely inspired by the silenced mating-type locus of *Schizosaccharomyces pombe* (Figure 2.9 for the basic model ingredients and selected results).

The main findings were:

- The model is able to produce high stabilities and heritability of either silenced or active states (bistability, Figure 2.9 C).
- Histone PTM domains can be highly dynamic without compromising overall stability, suggesting

that individual histone PTMs do not have to be stable permanently (Figure 2.9 B).

- Effective bistability of a system requires cooperativity in the positive feedback loops (Figure 2.10).
- Long range interactions (i.e., beyond nearest-neighbor nucleosomes) are advantageous because the system is otherwise very sensitive to stochastic fluctuations.

In follow-up papers, Dodd and Sneppen developed extensions of this model. For example, they investigated what effect TFs have in such a system [37]. The authors found that ultrasensitivity (Figure 2.10 B) can be achieved by TF binding, and that bistable promoter responses are possible by TFs that bind non-cooperatively to a single promoter.

In another variation of the model of Dodd et al. [36], Dodd et al. [38] added DNA elements, such as barriers and silencers[8], into the model to address the issue of potentially uncontrollable spreading of particular chromatin states. They found that:

- few histone-modifying enzymes and "simple" barriers (e.g., DNA-bound proteins or nucleosome-free DNA) may already suffice to obtain a system that (i) can be stably controlled in time and space and (ii) is ultrasensitive and bistable.
- not all reactions in the system must be cooperative for effective bistability, and having only one long-range positive feedback reaction with cooperative behavior may be already sufficient
- local silencer elements that exclusively act only on neighboring nucleosomes may control the state of the silenced region

Finally, Sneppen et al. [39] extended the original Dodd et al. [36] model to simulate the dynamics for two distinct histone PTMs rather than just one as modeled in previous approaches, each of which may be in an unmodified or modified state. Due to the complexity of the methods and results, see Sneppen et al. [39] for details. Briefly, the authors found that:

- even a small extension in the number of histone PTM states in combination with histone-modifying enzymes able to specifically recognize combinations of these two histone PTMs (histone code, see Section 2.1.4.1) creates large numbers of different modification and enzyme recruitment schemes that establish heritable bistability
- most bistable schemes were constructed so that nucleosomes in opposing states (i.e., `M` and `A`, in analogy to the model of Dodd et al. [36]) have different states at both histone PTMs
- all schemes include positive feedback and cooperativity between nucleosomes, and the system has multiple possibilities to generate cooperativity in the positive feedback reactions

[8] a DNA sequence capable of binding repressors (i.e., a transcription regulation factor that inhibits the expression of its target gene or genes)

- from an evolutionary perspective, such a system may easily evolve because most bistable schemes can easily be converted into each other due to their similarity while retaining bistability[9]

Angel et al. [280] also extended the model of Dodd et al. [36] to study vernalization. After fitting model parameters to experimental ChIP-Seq (chromatin immunoprecipitation followed by sequencing) data, the model predicted a bistable gene expression pattern of the floral repressor *FLC* in individual cells, which the authors indeed verified experimentally, thus confirming the suitability and usefulness of the recruitment-copying model.

Based on the experimental data of Hathaway et al. [277] (see Section 2.3.2.2), Hodges et al. [288] employed a mathematical model to study the spatial and temporal dynamics of H3K9me3 domains. They showed that a model that includes three processes (nucleation, propagation, and turnover; Figure 2.12) can reconstitute virtually all non-centromeric H3K9me3 domains in mouse embryonic stem cells and explain why H3K9me3 domains seem to be inherently bounded to a length of ≈10 kb. Thus, contrary to the model of Dodd et al. [38], it does not require boundary or insulator elements. Additionally, results suggest that inherently bounded domains arise if propagation occurs primarily locally (i.e., neighboring nucleosomes) instead of non-locally (i.e., beyond nearest-neighbor nucleosomes) although the latter may also produce them [288]. Hodges et al. [288] also found that the observed domain lengths are only compatible with their model if the relative propagation rate κ (Figure 2.12 D) does not exceed a value of ≈ 1.5. For higher values of κ, H3K9me3 spreads without bounds.

Analytical approaches confirmed the general requirements for bistability and cooperativity [30]. They also addressed additional biological phenomena such as how perturbations in the activity of histone-modifying enzymes affect the stability of epigenetic states [35] and what effect spatial dependence and spatially heterogeneous enzymatic activity has in such systems [31].

[9] for this, all bistable circuits were arranged as nodes in a network in which two schemes are connected if they have only one reaction catalyzed differently; see Figure 7 in Sneppen et al. [39]

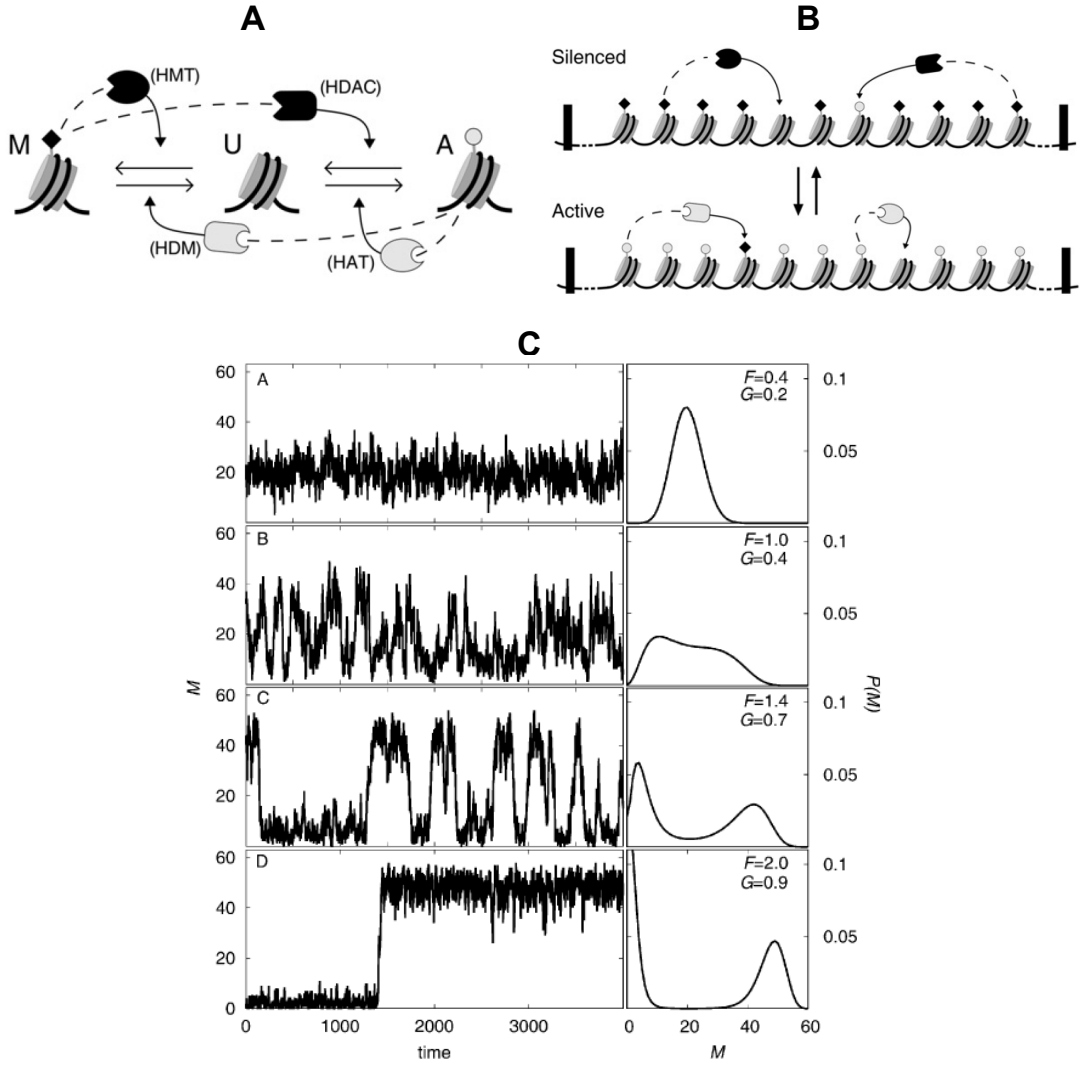


Figure 2.9: *The stochastic model for dynamic nucleosome modification from Dodd et al. [36]. The figure shows the basic model ingredients and selected results. For more details, see text. Figure taken from Dodd et al. [36].*
A: The model considers three distinct nucleosome states: `U` (unmodified), and the opposing states `M` (methylated, marked by a black diamond) and `A` (acetylated, marked by a gray circle). States can be interconverted by the recruitment of histone-modifying enzymes (names are abbreviated, see Section 2.1.4.4 for details) to nearby nucleosomes that are in the `M` or `A` state (dotted lines) or by random transitions.
B: In the silenced state, nucleosomes are predominantly methylated (`M`), whereas in the active state, nucleosomes are predominantly acetylated (`A`). Each nucleosome can principally stimulate the modification of any other [36]. Boundary elements are marked with black rectangles.
C: Illustration of the bistability of the system and its dependence on the feedback-to-noise ratio F using a system with 60 nucleosomes. Left: Samples of the time development (measured as average attempted conversions per nucleosome) of the number of nucleosomes M over a range of values for F (with F and therefore also bistability increasing from top to bottom). Right: Corresponding probability distribution of M obtained from long simulations.

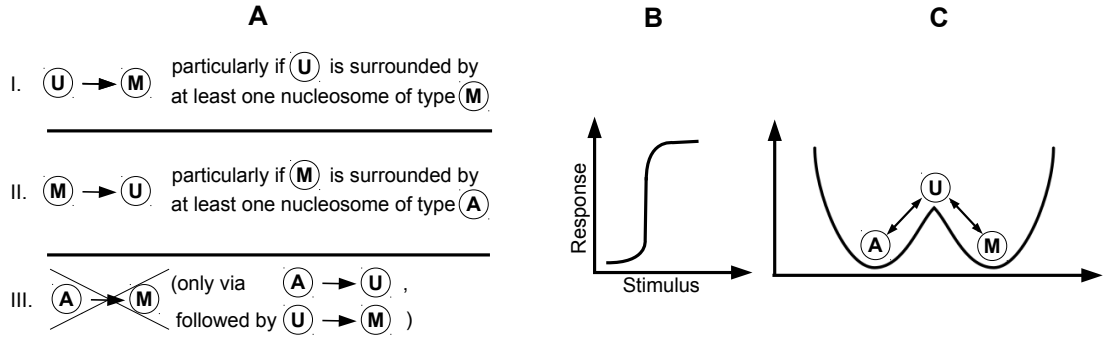


Figure 2.10: *Phenomena and properties of biological systems that are of relevance for epigenetic inheritance models. Three phenomena and properties are introduced, namely cooperativity in the feedback loop (A), bistability (B), and ultrasensitivity (C). B and C are modified from Wikimedia Commons files ("Bistability.svg" and "Ultrasensitivity.png", respectively), both of which are licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.*
A: Cooperativity is a "phenomenon of universal importance in biological systems and has almost as much variety as it has ubiquity" [281, p. 46841]. In enzymology, enzymes with multiple binding sites, each of which positively or negatively affects the affinity of the other binding site upon binding of a ligand by triggering an intramolecular conformational change, display cooperativity. In the epigenetic inheritance models as discussed here, the activity of more than one nucleosome in the modification reactions (i.e., reactions do not only depend on the state of the nucleosome that is subject to change but also on the states of additional nucleosomes) introduces cooperativity, for example by stimulating the addition of the same histone PTM on nearby nucleosomes (`I.`) or the removal of competing modifications (`II.`) [36]. The transition between opposing states also introduces cooperativity (e.g., `A` and `M`) requires two consecutive reactions (`III.`). Cooperativity therefore introduces a non-linearity in the feedback loop.
B: Ultrasensitivity (threshold behavior) is omnipresent in biological systems and particularly important for signal transduction [282]. It can be defined as a property of a system where "small changes in a stimulus near some threshold value produce a large change in a response but large changes in the stimulus far from the threshold produce small changes in the response (i.e., a sigmoidal stimulus-response curve)" [37, p. 1]. Ultrasensitivity arises from non-linear feedback mechanisms such as cooperativity.
C: Bistability is fundamental for many biological systems and may be defined as a system that exhibits two stable steady states that are separated by an unstable state [282]. Over time, the system may alternate between these two mutually exclusive and often opposing states. Ultrasensitivity in the feedback loops is a prerequisite for bistability [282, 283]. In models of epigenetic inheritance, these two states often correspond to active/acetylated (`A`) and methylated/silenced (`M`) regions, respectively, whereas the unmodified state (`U`) is only a non-stable transition state. Bistability requires both positive feedback and some kind of non-linearity in the feedback loop [284–286]. As stated in B, such non-linearity may arise through cooperativity in the feedback reactions, which, however, is not the only possibility (reviewed in [287]).

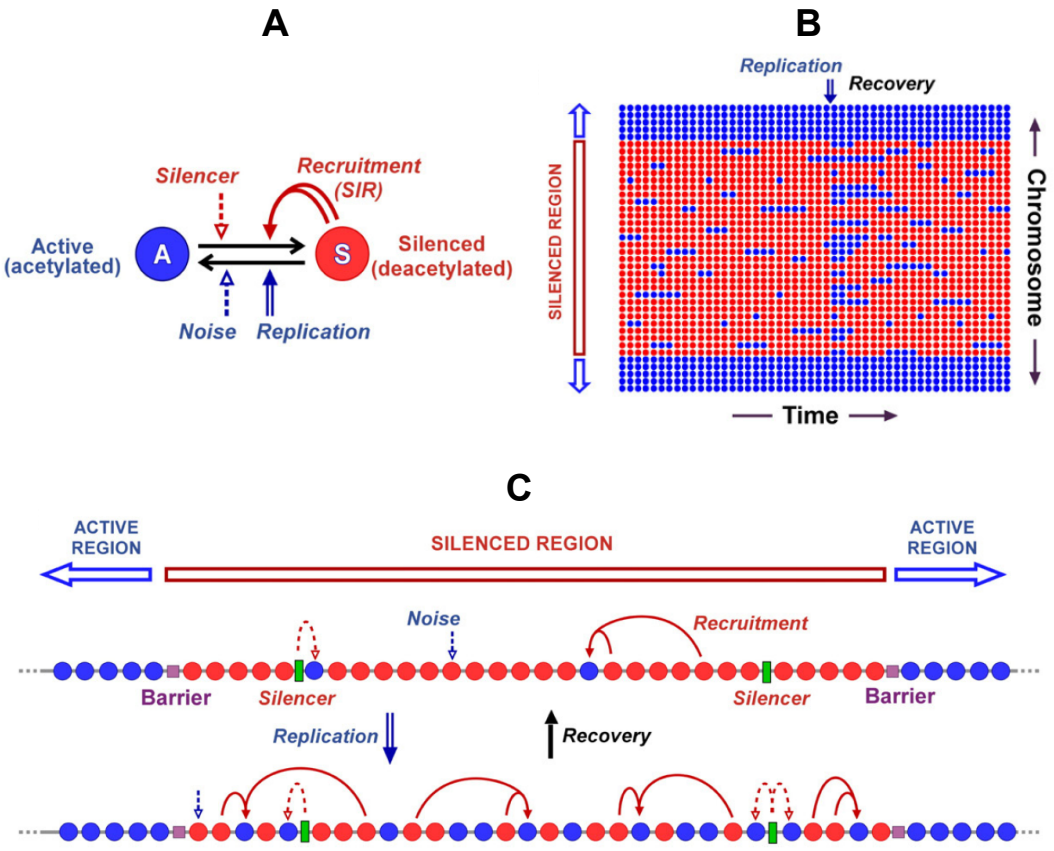


Figure 2.11: *The stochastic model for dynamic nucleosome modification from Dodd et al. [38]. Basic model ingredients are shown (see text for more details). Figure taken from Dodd et al. [38].*
A: The model considers the two opposing states `A` *(acetylated) and* `S` *(silenced). Note that the* `S` *state corresponds to the* `M` *state in Dodd et al. [36], whereas the* `U` *state is not part of the model. In analogy to the model of Dodd et al. [36], states can be interconverted by the recruitment of histone-modifying enzymes to nearby nucleosomes that are in the* `A` *or* `S` *state or by random transitions (noise). Each DNA replication converts the* `S` *state to* `A`*.*
B: Space-time plots of the evolution of a simulated system for the silenced mating-type locus (see C) consisting of 30 nucleosomes to illustrate the dynamics and mode of action of the model. Note that the authors displayed the nucleosome string vertically. In the middle of the simulation, the authors performed one DNA replication event, resulting in each nucleosome having a 50% probability of being replaced with a new nucleosome in state `A`*.*
C: Idealized structure of the silenced mating-type locus. The locus consists predominantly of a silenced region (`S`*). Barriers (depicted in purple) separate the silenced region from the surrounding* `A` *region. Silencer elements (depicted in green), defined as particular DNA sequences able to recruit the* SIR *protein (a HDAC that can modify neighboring nucleosomes), achieve maintenance of the* `S` *region. Additionally, positive feedback reactions further stabilize the* `S` *region by recruiting the* SIR *complex to nucleosomes in the* `S` *state to modify adjacent nucleosomes. In analogy to the model of Dodd et al. [36], conversion of the* `S` *state to* `A` *may occur via random, noisy reactions or DNA replication.*

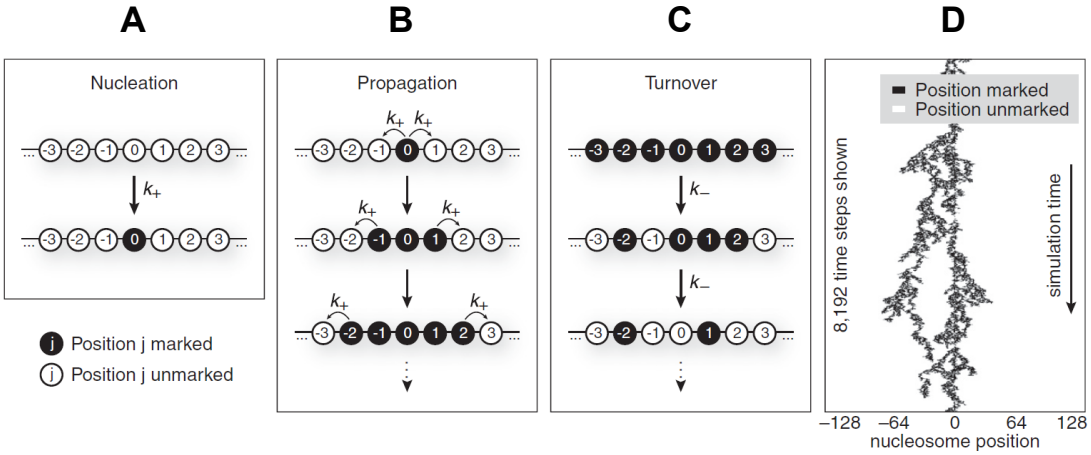


Figure 2.12: *The inherently bounded model of histone modification dynamics from Hathaway et al. [277]. The figure shows the basic model ingredients. According to Hathaway et al. [277], inherently bounded histone steady-state modification domains arise from the simultaneous action of three separate processes (A–C), and D shows an exemplary simulation iterating these three processes. For more details, see text. Figure taken from Hathaway et al. [277].*
A: Nucleation: Unmodified target sites are modified with a probability of k_+
B: Propagation: Unmodified neighboring nucleosomes are unmodified with a probability of k_+.
C: Turnover: For each modified nucleosome, conversion to an unmodified state occurs with probability k_-.
D: Exemplary simulation consisting of 8,192 time steps that iterates the processes in A–C to produce an inherently bounded domain. The relative propagation rate $\kappa = \frac{k_+}{k_-} = 1.5$. *In this example, domains transiently bifurcate to form multiple smaller domains.*

In addition to the recruitment-copying model, researchers analyzed other epigenetic inheritance phenomena either analytically or mathematically. Based on quantitative mass spectrometry measurements for budding yeast, for example, De Vos et al. [170] specifically addressed whether H3K79me qualifies as an epigenetic mark. They used an analytical model as well as *in vivo* experiments to study the propagation of H3K79me by the non-processive enzyme *Dot1* throughout cell division and found that H3K79me never reaches a steady state because H3K79me1, H3K79me2, and H3K79me3 all differ in their kinetics [170, 289]. Instead, H3K79me accumulates over time, only counteracted by the replacement of parental histones. Importantly, the authors observed the slow accumulation of all levels of H3K79me during the cell cycle for both new and old (i.e., parental) histones. Thus, H3K79 methylation represents a timer for histone age and couples cell-cycle length to chromatin changes.

Lastly, Radman-Livaja et al. [260] investigated if parental histones reassociate in the vicinity of their original locus from which they were evicted. The authors developed a mathematical model to explain their experimental results. Although not directly a model for epigenetic inheritance, retainment of parental histones after DNA replication is of relevance for the recruitment-copying model, as pointed out in Section 2.3.2.2. For more details and results, see Section 2.2.5.

Part I:

EUKARYOTIC CHROMATIN AS A MOLECULAR COMPUTER

Chapter 3

The Cellular Chromatin Computer

3.1 Motivation and Background

3.1.1 Standard and Non-Standard Computation

Computation is paramount to the discipline of computer science, yet a solid and universally accepted definition is lacking because it is difficult to define precisely what it means to "compute" [290]. Consequently, computation has been intensively discussed and characterized in many different ways [290–294]. A traditional and well-known definition stems from the Church-Turing thesis, which, in brief, states that a function is algorithmically computable if and only if it is computable by a Turing Machine (TM) [295]. Thus, computation encompasses everything that can be simulated by a TM. This interpretation of the Church-Turing thesis continues to be the one most widely adopted by computer scientists [290]. The TM is still the most commonly examined model of computation.

A more detailed discussion of how computation may be defined and what it encompasses is out of the scope of this thesis. However, in its most general sense, computation may simply be regarded as information processing. Information processing itself may be defined as a series of three step sequences consisting of acquiring information from the outside world (input), transforming it in a particular way (data manipulation), and finally providing the outcome to the outside world (output) [290]. Both human-initiated and naturally occurring computations apply equally well to this general definition. However, the question whether computation and information processing are identical concepts has been discussed repeatedly by various authors [294, 296, 297]. Although both terms are often used interchangeably and frequently indeed mean the same, some authors disagree that they should be used synonymously in general [294, 296].

Despite its exact definition, computation is nowadays typically immediately associated with the traditional silicon-based computer technology that is omnipresent in our daily lives. However, more and more information processing and computation are discovered as fundamental processes of many fields [3].

These non-standard (also called unconventional or alternative) computing paradigms include, for example, the following technologies and disciplines:

- ternary computing (using a ternary instead of the common binary logic)
- quantum computing (using the principles of quantum physics and quantum-mechanical phenomena as operations such as superposition and entanglement)
- optical computing (using light instead of electricity for computation)
- natural computing (computational processes that are observed in nature)

Natural computing is particularly diverse and may furthermore be divided into the following fields:

- molecular computing[1] (using chemical molecules such as DNA, RNA and proteins for computational tasks)
- evolutionary computing (computation based on fundamental principles of biological evolution such as natural selection, recombination, and mutation)
- membrane computing (computation inspired from the structure and the mode of action of living cells and their organization in higher order structures [298])
- neural computing (computational models that are inspired by the central nervous system and the brain in particular)

Intriguingly, some phenomena and models in natural computing are based on methods of formal language theory such as development of multicellular organisms (L systems [299]), cellular automata (see Wolfram [300] and references therein), membrane computing (P systems [298]), and DNA computing (splicing systems or H systems [301]). Particularly the work from Head [301] was highly influential. He investigated the effect of the biomolecular operations of restriction enzymes from a computational perspective, which subsequently stimulated the design of DNA computers (see next section).

3.1.2 Natural Computing and the Role and Significance of Chromatin in the Cellular Computation Machinery

Natural computing is a particularly fast-emerging and fascinating area because it originates from naturally occurring biological systems. Ever since their initial discovery, natural computing techniques inspired the development of novel problem-solving techniques, for example by evolutionary optimization and swarm intelligence algorithms. Intriguingly, it becomes increasingly clear that computing may not only be regarded as an artificial science but also as a natural one [3–5].
One somewhat exotic example for natural computation is membrane computing. It does not denote

[1] also called genomic [7], biochemical, natural or cellular computing

a specific model or theory; instead, it must be seen as a general framework that pays particular attention to membranes and compartmentalization and consequently also to various other related concepts such as communication, distribution, localization, and hierarchization [302].

A more well-known example is molecular computing, which tries to both use molecules for computation and understand the information processing and computational nature of molecular processes in general. Cells, the "building blocks of life", are incredibly complex and highly sophisticated biological units with huge information processing capabilities. Researchers therefore regarded the cell or specific cellular components repeatedly and with increasing frequency as a cellular computer capable of performing complex biological (*in vivo*) "computations" [4, 6–14]. Indeed, advanced information processing with molecules inside living cells is omnipresent and occurs on all scales. It can be found, for example, in complex structures such as the brain [15], in regulatory and signaling pathways within cells, or even within single biomolecules [16].
In molecular computers, DNA is typically exploited as programmable matter to perform computation and therefore exemplifies the information carrier and storage molecule, whereas its products — proteins and RNAs — react to external stimuli and perform complex molecular tasks that ultimately govern and control what occurs within the cell.

Two specific examples highlight the capability of cells to perform computation. First, in 1994, Adleman et al. [303] demonstrated that computationally hard problems, such as specific instances of the Hamiltonian path problem, can be solved by manipulating DNA strands [303]. Briefly, the Hamiltonian path problem determines whether a Hamiltonian path exists in a given graph with a set of vertices and edges. It belongs to the class of NP-complete problems. A Hamiltonian path is a path that passes through every vertex in a particular graph exactly once[2]. Adleman et al. [303] used a non-deterministic algorithm to solve the Hamiltonian path problem using a "DNA computer". It was the first real example of a cellular computer that can perform an actual calculation and therefore denoted a milestone for synthetic biology. Due to the importance of this work and as illustration of how a natural computer may function, the basic principles and mode of action of his DNA computer are depicted in Figure 3.1. Another noteworthy demonstration of the computational capabilities of DNA comes from Siuti et al. [304] who showed that Boolean logic functions may be implemented solely with stable DNA-encoded memory.

[2]Note that a Hamiltonian path does not necessarily also pass through all the edges of the graph.

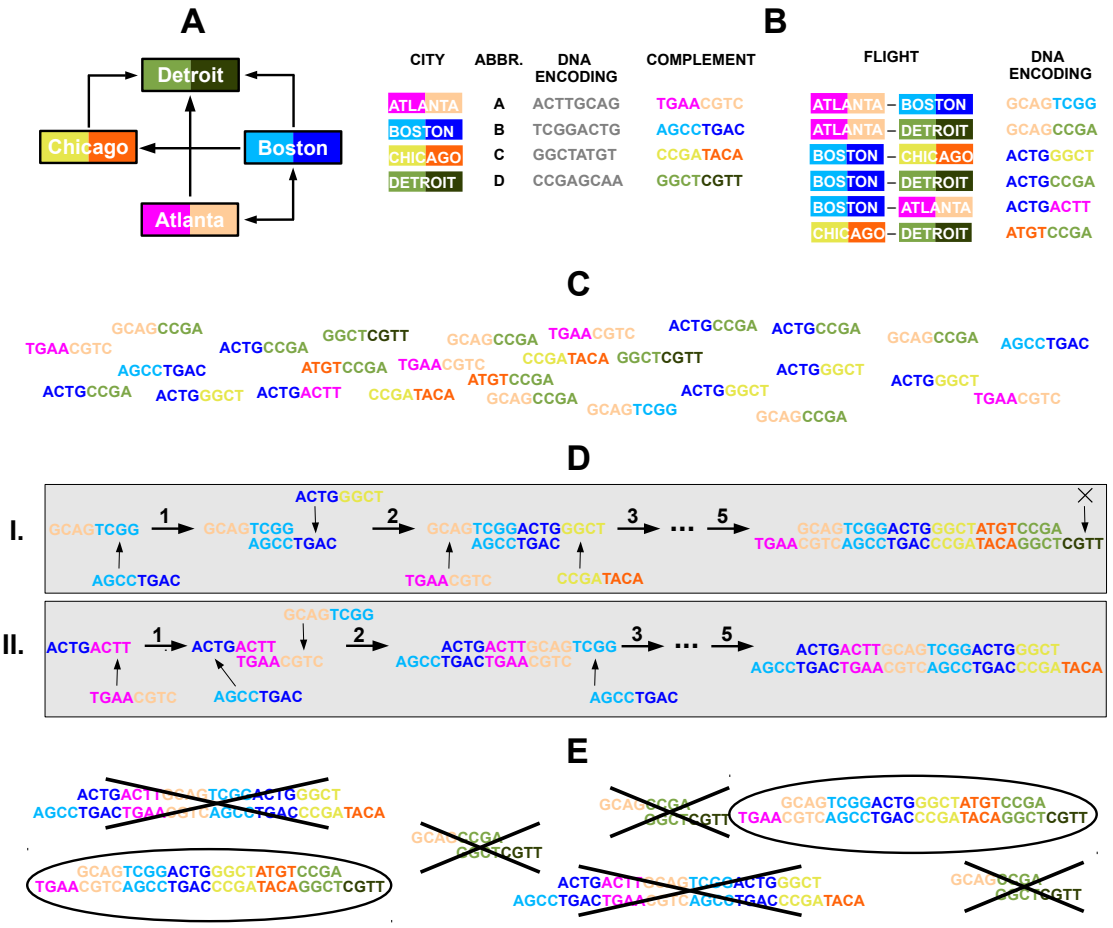


Figure 3.1: *Example DNA computer for the Hamiltonian path problem. The DNA computer solution is based on the following non-deterministic algorithm: 1. Generate random paths; 2. Keep only those paths that fulfill all criteria for a Hamiltonian path; and 3. If paths remain, a Hamiltonian path exists. The graph and the DNA encoding is taken and modified from Adleman [305].*
A: An example graph that represents a map of four cities (vertices) that are connected by various nonstop flights (edges). For this graph, a Hamiltonian path exists (A – B – C – D).
B: Each vertex and each edge in the graph is assigned a particular single-stranded DNA sequence. DNA encodings generally consist of two parts (see coloring). For each city, the complementary sequence of the DNA encoding (denoted as $\overline{S_{city}}$) and its abbreviation is also given. DNA flight encodings (S_{flight}) are constructed by concatenating the last four nucleotides of the city of origin with the first four nucleotides of the city of destination of the corresponding DNA encodings.
C: Generation of many copies of sequences of the types $\overline{S_{city}}$ and S_{flight}. The sequences are synthesized in large numbers and then brought together in a tube.
D: Generation of random paths (step 1). Random paths are generated in a ligation reaction by mixing the DNA strands with DNA ligase and adenosine triphosphate in a highly parallelized manner. DNA strands with complementary sticky ends may therefore ligate. The stepwise construction of two example paths is shown: I. A–B–C–D (Hamiltonian path) and II. B–A–B–C.
E: Elimination of solutions that do not represent Hamiltonian paths and determination whether a Hamiltonian path exists (steps 2 and 3). Filtering occurs stepwise and is based on various biochemical tools such as gel electrophoresis and PCR amplification (step 2). All paths that remain represent a solution to the problem (step 3). Due to lack of space, I refer to Adleman et al. [303] for more details (in particular how the various filtering steps are constructed).

The second example for the capability of cells to perform computation comes from the gene regulatory system. Various authors analyzed the gene regulatory system and particularly *cis*-regulatory modules (CRMs) intensively in the last decade for their computational potential [7, 8]. CRMs are abundant ($>$ 100,000), mostly relatively short ($<$ 1 kb) stretches of DNA that can be bound by multiple TFs. Upon binding, CRMs regulate the expression of nearby genes by interacting with the transcription machinery (regulation *in cis*). One CRM may regulate several genes but one gene may also be controlled by various CRMs. CRMs frequently have multiple TF binding sites (inputs), and the output they produce depends on the combinatorial absence or presence of all input molecules. Intriguingly, their regulatory outputs may often be derived by applying the basic Boolean functions `AND`, `OR`, and `NOT` to the specific input pattern. However, the Boolean logic operations are only a suitable approximation for particular instances and not a general principle for CRMs [306]. Nevertheless, CRMs are capable of performing computational logic operations. Input molecules themselves as well as down-stream effectors may be independently regulated by CRMs [12], thus further increasing computational complexity. CRMs therefore may function as information processing devices because they are interconnected and form large regulatory networks that control, for example, organism development [7]. Istrail et al. [7] consequently termed the collection of all CRMs a "genomic computer" due to their capability to produce complex circuits and compared mode of action of CRMs and ordinary, man-made computers.

However, not only TFs serve as input for CRMs. Chromatin-associated phenomena, such as histone PTMs, histone variants, DNA methylation, and the specific chromatin structure (e.g., higher-order organization and nucleosome spacing), also determine the functional output of CRMs due to their tight coupling to transcription and gene regulation more generally (see Chapter 2). TFs may not be able to bind to their designated target sites (that is, CRMs) due to the wrapped structure of the region in the presence of nucleosomes, for example (see Section 2.1.2). Thus, regulation on the chromatin level is intermingled with CRM functionality but most likely also represents an independent and additional regulatory circuit. Prohaska et al. [12, p. 37] even argued that "chromatin regulation adds a computational layer that, in Eukarya, is qualitatively different and potentially more powerful than the CRM networks". The part of the cellular computation machinery associated with chromatin has remained largely unexplored so far although various authors recently postulated repeatedly that chromatin may act as a computational device capable of performing "computations" in a biological context [12, 13, 17]. Indeed, as shown in Chapter 4, the reconstitution of local patterns of histone PTMs after DNA replication is one example of the biologically important computational tasks that can naturally be solved by the "chromatin computer" (CC).

3.1.3 Evolution of Chromatin and Chromatin-Based Regulation

To appreciate the computational capabilities of chromatin and to illustrate why computational capabilities and memory capacities beyond *cis*-regulatory networks are useful for gene regulation, it

is important to realize how chromatin and chromatin-based regulation in the cell evolved. In what follows, a summary of the evolution of chromatin is presented that is mainly based on an inspiring publication by Prohaska et al. [12] who employed a phylogenetic comparison of the chromatin-based regulatory system among the three domains *Eubacteria*, *Archaea* and *Eukarya*.

CRMs evolved step-wise and therefore comprise a mixture of regulatory structures of different ages [7]. Similarly, the chromatin regulatory system evolved step-wise through a number of key molecular inventions that substantially expanded the cell's regulatory scope and its computational power [12] (Figure 3.2). Throughout evolution, the structure and organization of chromatin also became more complex and diverse. For example, nucleosomes and histones are only found in *Archaea* and *Eukarya* but not *Eubacteria*, whereas histone tails are only present in *Eukarya*.

Prohaska et al. found that mode of operation and complexity of chromatin-based regulation differs substantially among the three domains (Figure 3.2). Initially, chromatin was a very general and crude transcription regulator [12]. Ancestral ways of regulation include the destabilization of chromatin by the incorporation of variants of chromosomal architectural proteins. Whereas this mechanism is present and used in all three domains of life, it is particularly dominant in *Eubacteria*. To extend the regulatory scope of the cell, chromosomal architectural proteins may be modulated either structurally or by altering their binding properties. Whereas chemical modifications are very rare to non-existent in *Eubacteria*, they are widespread in both *Archaea* and *Eukarya*. Such modulations seem to be more flexible and resource-saving than the expression of different paralogs of particular chromosomal architectural proteins as in *Eubacteria*. The ability to chemically modify them was thus a major innovation in the evolution of chromatin.

In *Eukarya*, finally, a second significant component emerged: modification readers that can specifically recognize particular (combinations of) histone PTMs. This transition from a write-delete to an elaborate read-write-delete system resulted in a substantial increase in computational complexity, memory capacity, and regulatory flexibility. It turned chromatin into a cellular memory device able to keep a record of former metabolic states such as local transcriptional activity or DNA damage. It therefore does not seem coincidental that different co-occurrences of reader and modifier (writer or eraser) domains are increasingly common and diverse in *Eukarya* in general and particularly in *Metazoa* and that the increase in computational complexity correlates well with the emergence of complex multicellular life. The latter observation seems plausible because cell differentiation and the subsequent retention of cellular identities seem to require chromatin-based and particularly epigenetic mechanisms. Indeed, epigenetic inheritance seems to be the latest addition to the chromatin-based regulatory system [12].

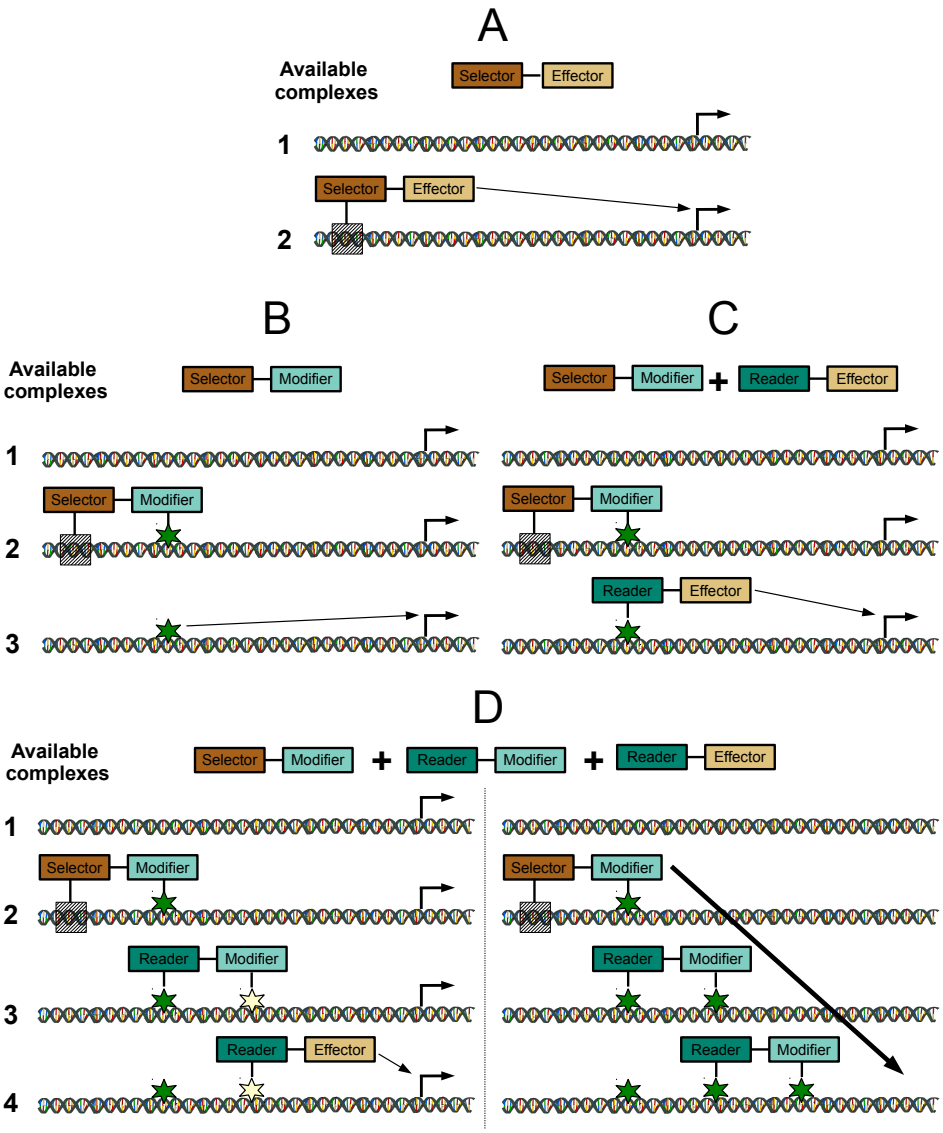


Figure 3.2: *Overview of conceptual innovations in the regulation of chromatin during evolution in Eubacteria, Archaea, and Eukarya. The regulatory power and complexity increases from A to D. Selectors bind to a specific DNA locus (e.g., DNA binding domain), effectors induce a particular event (e.g., a transcriptional change), modifiers induce a chemical modification, and readers can read or recognize such modifications. For more details, see text. Modified, with permission, after [12].*
A: Selector-effector pairs have a direct effect on transcription (present in Eubacteria, Archaea and Eukarya but the dominant regulatory principle in Eubacteria).
B: Selector-modifier pairs can covalently modify chromosomal architectural proteins (only Archaea and Eukarya). Once the modification is set, a permanent effect is carried out.
C: Selector-modifier and reader-effector pairs allow to execute an effect independently of the original establishment of the signal (only higher Eukarya). The modification now serves as a memory or a signal in an information theoretic sense and can be interpreted differently by reader-effector pairs.
D: Selector-modifier, reader-modifier and reader-effector pairs finally allow even more complex computations (left) and signal propagation as well as signal inheritance (right) (only higher Eukarya). Note that using this mode of regulation, there may not even be an immediate regulatory effect.

3.2 Methods and Results

3.2.1 Components

Ordinary silicon-based computers are composed of the four components input and output device, memory, execution unit, and control unit [9]. To a greater or lesser extent, this holds also true for the cellular computer, and Istrail et al. [7] and Moe-Behrens [9], for example, provided mappings from cellular to ordinary computers. Whereas Istrail et al. [7] specifically focused on CRMs as reference for a cellular computer, Moe-Behrens [9] more generally mapped various biological building blocks to system components of general purpose computers. In what follows, I therefore focus on parts specific to the CC that were not yet adequately described and similarly map them to the four general system units.

3.2.1.1 Memory

Memory is a fundamental component of each computer. Similarly, particularly for more complex organisms, it is crucial to keep a molecular memory of past stimuli and states. Such memories can come in many facets and include, for example, a transcriptional and cell identity memory [64, 185], DNA damage memory [166, 307], immunologic memory [308, 309], metabolic memory [310] and a memory of prior stress exposure [311, 312]. More generally, these kinds of memories are frequently referred to as epigenetic memory. Intriguingly, it has become increasingly apparent that chromatin plays a significant role in all these examples, suggesting that chromatin provides enough flexibility to store both transient and more permanent states (signals).

In the CC, the fundamental memory unit is the nucleosome or more specifically the histone octamer, which represents a highly versatile memory page (Figure 3.3). Memory as a whole is therefore constituted by the collection of all nucleosomes. Although the depiction of nucleosomes arranged as a linear chain in Figure 3.3 may be generally a reasonable approximation, in reality, chromatin forms higher-order structures (see Section 2.1.4.2). Thus, pairs of nucleosomes that are in large spatial distance from one another when arranged as a linear chain may in fact be in close proximity due to the three-dimensional organization of chromatin.

Similar to ordinary computers that typically consist of multiple kinds of memory (e.g., RAM and hard disks), the CC also has additional memory capabilities such as the various forms of DNA methylation to store both transient and more permanent information. In analogy to the histone proteins, chromatin-modifying enzymes and more generally all proteins or RNAs that directly or indirectly interact with chromatin (e.g., *Polycomb* and *Trithorax* group complexes) may be subject to a wide array of PTMs, all of which have a particular meaning that may alter the specific "program" that is executed. Even the particular higher-order chromatin structure or the absence of nucleosomes

at particular positions may be regarded as a particular kind of memory because both typically resemble the transcriptional status of the genomic region that was established previously.

The CC therefore predominantly utilizes a passive memory (that is, information is written once and subsequently interpreted by the cell until erased or overwritten [7]) in the form of histone and non-histone PTMs and DNA methylation. Pure CRMs do not have such an extensive passive memory facility and use two forms of an active one instead. First, they maintain their transcriptional state by continuously activating particular transcriptional subcircuits and therefore expressing genes and corresponding downstream targets through intracellular positive feedback loops. Second, they exhibit the "community effect"[3] through intercellular positive feedback circuits [7]. Whereas it may be sufficient for simple organisms to maintain transcriptional states solely by self-propagating feedback loops and *cis*-regulatory TF networks [64], a cellular memory as exemplified by higher organisms has the major advantage that it is no longer necessary to instantly activate or terminate transcription by means of metabolites or environmental signals [12].

In contrast to hard disks of ordinary computers, memory in biological systems is inherently non-permanent because the signals that represent the memory may be degraded, transformed, and generally altered. As outlined in Section 2.1.4.1, histone PTMs, for example, have strikingly different lifetimes. Acetylation events are measured in the order of minutes, whereas histone methylation are more stable. To some extent, this therefore has some analogies to RAM, which typically also loses its information in particular cases (i.e., when power is turned off). Although not part of the CC, TF networks may also be compared with RAM because the memory has to be repeatedly re-established in order to carry out its function.

3.2.1.2 Execution Unit

Arithmetic and logical operations are represented by chromatin-modifying enzymes, the "processors" of the CC. They also mediate the transitions between chromatin states. Chromatin-modifying enzymes work largely independently from one another and catalyze the often context-dependent writing or erasing of histone PTMs and thus directly modify the memory of the CC (Figure 3.4). For context-specificity, they often also contain reader domains that can specifically recognize the presence or absence of one or multiple histone PTMs (see Section 2.1.4.4). Indeed, chromatin-modifying enzymes implement logical operations through their reader domains. The combinatorial recognition of distinct histone PTMs (histone crosstalk) already implements logical operations analogous to `AND`, `OR`, or `XOR` gates in digital circuits. Two reader domains or even a single one are sufficient for selectively recognizing multiple histone PTMs (see Section 2.1.4.4). These logical operations may be represented as read-write rules. Due to the presence of multiple reader domains, the semantic complexity of these rules can be arbitrarily high although little is known about the

[3] intra-territorial, mutual signaling among cells of a particular territory to retain a similar transcriptional state [7, 313]

true biological complexity for large multisubunit complexes. Noteworthy, the available players in the chromatin regulatory system (Figure 3.2) represent various logic control structures that have a programming analogy. Selectors, for example, logically correspond to an "if" statement, whereas a modification in *Archaea* implements a "while" statement because the effect is carried out as long as the modification is present. The combination of these elements enables almost arbitrarily complex operations and therefore a very fine-tuned regulation (particularly in higher eukaryotes).

From a theoretical point of view, the collection of rules may be regarded as a term rewriting system (see Chapter 4 for specific rule examples). Term rewrite systems are sets of directed equations with a particular simple syntax and semantics [314]. They replace certain patterns (i.e., the left side of the rule) in terms by other terms (i.e., the right side) [315] and are standard for handling strings in formal language theory. Generally, the right sight may sometimes be composed "simpler" although this is not the case with the rules in the CC (see Section 4.2.2). Term rewrite systems are Turing-complete and therefore conceptually very powerful models of computation (see Section 3.2.4) (reviewed in [314]). They may therefore be used for computation because they provide a fully general non-deterministic programming language [314]. For example, they may be used to execute logic programs [314]
The addition of particular constraints how these rules are constructed may result in substantially simpler modes of computation. For example, if variables are not allowed in the rules, one obtains ground term rewrite systems. Another restricted kind of a term rewriting system is a string rewriting system (also called a semi-Thue system), a term rewriting system over strings from a usually finite alphabet. The rules that I construct in Chapter 4 (see Section 4.2.2 in particular) fulfill the simplifying requirements of string rewriting systems although construction of more complex rules that contain variables is easily imaginable and maybe even necessary for realistic modeling of the underlying biology of chromatin-modifying enzymes.

lncRNAs are also important for the logical operations of the CC by guiding chromatin-modifying complexes to particular genomic loci and therefore contributing to the regulation of gene expression (see Section 2.1.4.3). Rather than representing simple scaffolds, they may even represent complex "computer circuit boards" that link together various molecular components [316]. A CC therefore has its own logic, independent of the logic from pure CRMs.

Histone-modifying enzymes may just evaluate the state of the histone residue(s) that they modify, independent of the modification state of other histone residues on the same or neighboring nucleosomes, or they may depend on the states of other histone residues (context-independent and context-dependent rewriting rules, respectively, see Figure 3.4). As reviewed in Section 2.1.4.4, context-dependent rewriting rules indeed seem to be common for eukaryotic systems.
The (relative) frequencies of rule execution may simply be modulated by altering the concentrations of the corresponding histone-modifying enzymes (see Chapter 4).

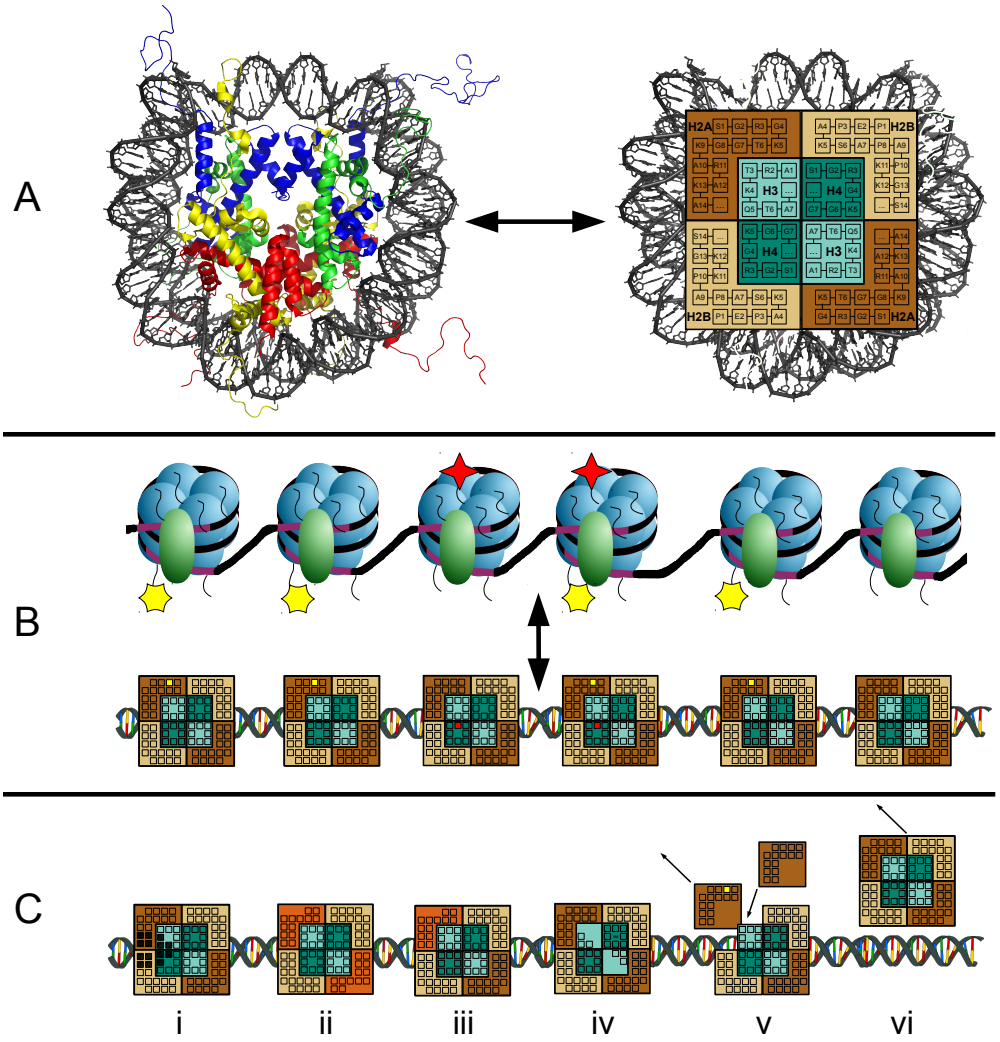


Figure 3.3: *The nucleosome as a versatile memory page.*
A: Crystal structure of a eukaryotic nucleosome with all four core histones (left) and its representation as a memory page (right). The memory page is an abstraction and integrates all histones, each of which is composed of those amino acid residues that may carry a PTM (Figure 3.5). For illustration purposes, however, the beginning of the N-terminal domain is shown regardless of whether the residue may be in at least two different states (based on human histones, shown as single letter codes). Note that the linker histone H1 is not part of a nucleosome and therefore not included here.
B: Abstraction of chromatin as a linear sequence of memory pages, analogous to the well-known "beads-on-a-string" picture. Two distinct histone PTMs are highlighted (yellow and red, respectively), each of which with its own occurrence pattern.
C: Individual memory pages may dynamically change their structure, content and properties. Such versatile changes include (from left to right) the temporal non-availability of individual registers/memory cells (marked as dark shaded) due to their occupation with bound proteins, for example (i), the presence of histone variants in symmetric and asymmetric configurations (ii and iii, respectively), tail clipping of individual histones (iv), individual histone exchange with newly assembled and unmodified histones (v) or even modified histones (not shown), and eviction (depletion) of full nucleosomes (resulting in a nucleosome depleted region) (vi). For simplicity, in (vi), the nucleosome is depicted as a unit during eviction, whereas in reality it is decomposed into histone multimers or individual histones.

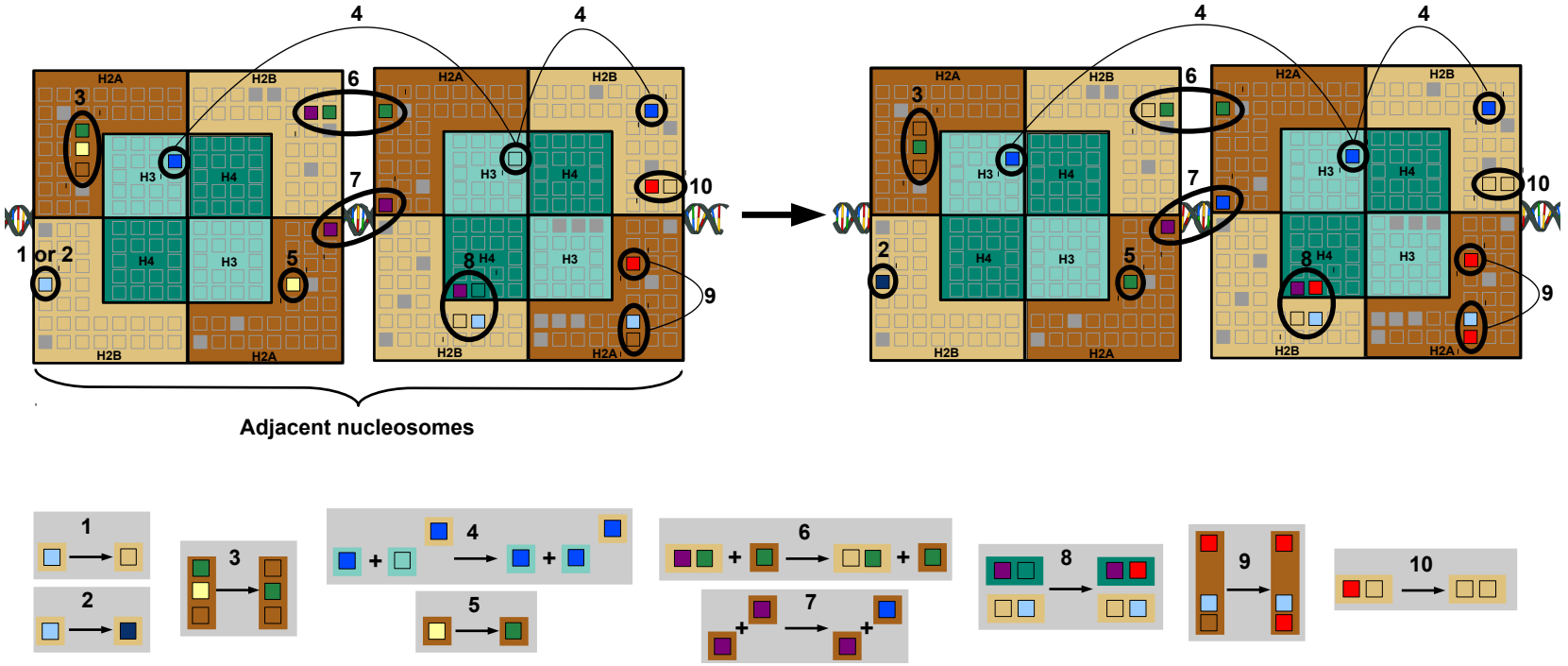


Figure 3.4: *Variety and complexity of the chromatin computer rules. Two (spatially) adjacent nucleosomes (abstracted as memory pages, see Figure 3.3) are shown and a total of ten reactions that may be performed by various chromatin-modifying enzymes. Each nucleosome is in a particular state, as indicated by the different histone PTMs, each of which is colored distinctively. Histone PTMs irrelevant for the highlighted reactions are colored in gray. Reactions can generally be context-independent (1, 2, and 5) or dependent on at least one other position/residue (all other rules). Rules may depend on the state of particular residues within the same histone (3, 8, 9, and 10), at different histones but within the same nucleosome (8), within the same histones but on a different nucleosome (7), at different histones on a different nucleosome (6). These four classes may also be combined arbitrarily (4, 6, and 8), particularly if rules take into account the modification states of more than two residues (3, 4, 6, and 8). Additionally, different rules may compete for the same amino acid residue (1 and 2), of which only one is executed (2).*

3.2.2 Mode of Operation and General Properties

In this section, I specifically focus on mode of operation, general properties, memory size, computational universality and efficiency, and speed of operation of the CC. I will also investigate which of the traditional computer architecture paradigms the CC has most similarities with. To the best of my knowledge, all of this has not been done systematically so far.

The process of writing and erasing histone PTMs is intrinsically stochastic and crucially depends on concentrations of the available histone-modifying enzymes. A CC consequently has a set of read-write rules that operate non-deterministically on chromatin (see Section 3.2.1.2). Chromatin itself can be abstracted as a string of nucleosomes, analogous to a TM (see Section A.2) on a finite tape. A cellular automata-like 1-D string as the computational paradigm for the CC is therefore proposed on which sets of local rewriting rules are applied asynchronously with time-dependent probabilities. The "software" for this type of CC is thus a sequence of sets of rewriting rules and concentrations that can be directly interpreted as part of the cell's gene expression program. However, rewriting rules are also the "hardware" because they represent physical entities that execute a particular function: chromatin-modifying enzymes. A distinction into hardware and software is therefore not useful for biological computers because they are physically inseparable [7].

A CC thus best belongs to the "Multiple Instruction, Multiple Data streams" (MIMD) architecture in Flynn's taxonomy [317] because multiple autonomous processors may simultaneously execute different instructions on different data. Furthermore, a CC provides a shared memory because all processors have access to a globally available memory.

Regarding computer architecture paradigms, a CC may be considered a register machine as opposed to a stack machine due to the presence of a large number of mostly uniquely addressable registers. Nucleosomes have relatively stable positions although phenomena such as nucleosome sliding may occur. However, if nucleosome positions were solely determined by DNA, then nucleosomes in repeat regions would be difficult to address uniquely due to the repetitive sequence in the vicinity of the nucleosome.

The CC combines uniform and non-uniform memory access. Part of the memory cells are easier to access than others due to different higher-order chromatin structures and accessibilities (e.g., heterochromatic versus euchromatic regions). However, access time within heterochromatic or euchromatic regions is likely to be approximately equal. It is in general also independent of the total size of the memory and directly accessible in any random order. Additionally, chromatin-modifying enzymes may have their own local, limited memory that can be accessed faster. For example, they may be modified in their structure, composition, and modification state, all of which represents information that can be interpreted by the cell. In general, chromatin memory is largely addressable and writable (e.g., histone PTMs and TFs that may specifically bind particular DNA sequences and therefore also nucleosomes, respectively). However, memory pages are

- variable with respect to their specific position (e.g., although each nucleosome typically belongs to a particular genomic locus, this position may change due to nucleosome sliding mechanisms [318])
- versatile (e.g., histone PTMs are partially reset after each DNA replication, see Section 2.2)
- heterogeneous (e.g., the structure of particular memory pages may be different, Figure 3.3)

Molecular computers, such as the CC, are principally asynchronous, because they are not governed by any global clock over short time scales. Chromatin-modifying enzymes operate independently and in a highly parallel manner. Indeed, such asynchrony provides many advantages as compared to synchronous systems [319, 320]. For example, the CC is tolerant against temporal variation and local asynchronies because temporal synchrony is replaced by causal dependencies among regulatory networks [7]. However, considering longer time scales, eukaryotic cells do contain two endogenous clocks that are interlocked to some extent: the circadian clock and the cell division clock [321]. Shaped by the day-night cycle, the circadian clock generator produces a timing signal with a periodicity of approximately 24 hours that regulate gene expression of a system of 'clock genes' [322, 323].
Intriguingly, the very same mechanisms that define the mode of action of the CC[4] play important but not yet fully elucidated roles for the establishment of cellular clocks [322, 323].

The CC has additional properties with striking analogies to amorphous computing[5][11, 324]:

- Self-organizational and emergent behavior because the global system pattern "emerges solely from numerous interactions among the lower-level components of the system. Moreover, the rules specifying interactions among the system's components are executed using only local information, without reference to the global pattern." [325, p. 8]
- Redundancy (e.g., various histone PTMs have identical or similar functions)
- potential presence of erroneous or faulty components that may lead to error-proneness (e.g., non-functional chromatin-modifying enzymes or misfolded histone proteins), which, however, can be tolerated up to a given threshold without malfunctions.
- Chromatin-modifying enzymes, the workhorses of the CC,
  - have no or very limited memory capacity and computational abilities
  - typically have no a priori knowledge of their specific location (more universal binding, see above)
  - act asynchronously and perform their designated reactions independently of one another

[4] i.e., histone PTMs such as histone acetylation and deacetylation and more generally alterations of the chromatin structure

[5] computational systems characterized by large numbers of irregularly placed, identical, and asynchronous computing elements of limited computational ability that predominantly interact locally [11]

- communicate mainly locally
- are present in large numbers (up to multiple millions of molecules [289]) and work in a highly parallel manner

Thus, similar as for biochemical computers in general, the mode of operation for a CC is conceptually similar to well-known concepts from the complex systems theory such as non-linear bifurcations; interlocking positive and negative feedback loops; distributed networks and information control; implicit and explicit cooperativity; redundancy; self-organizational (emergent) behavior; and context-dependency [30, 36, 38, 39, 326]. A CC may also modify its own program during computation. Such self-modification may be achieved, for example, by altering the level of transcription for the chromatin-modifying enzymes that are part of the computation or by regulating chromatin structure.

The number of operations per second can hardly be reliably estimated, for reasons outlined in Section 4.4. In particular, reported values for processes required to estimate the number of operations per seconds, such as average residence times and the number of molecules per cell, span several orders of magnitude and furthermore highly depend on the specific protein [289]. Bryant [13] nevertheless estimated the number of operations per second using two different approaches and yielded values of 10,000 and 1,000,000 (i.e., 0.01 MHz and 1 MHz), respectively. However, the former calculation was only based on data for RNA polymerase II and is therefore an underestimation, as also noted explicitly [13]. The latter estimate stems from the crude estimation that average read, write, and/or delete operations take 1 second. Although average values are principally unknown and *in vivo* kinetics for residence times vary widely among chromatin binding proteins [289], existing data indicate that for some proteins, they can indeed be in the range of approximately one second (e.g., human RNA polymerase I). The latter calculation also assumes that approximately 1% of all nucleosomes are bound by chromatin-modifying enzymes, which is, however, not biologically motivated and simply an educated guess.

3.2.3 Memory Size

Various authors suggested repeatedly that a CC may store and process more information than pure *cis*-regulatory networks [12, 13]. However, few approaches explicitly estimated the information content (IC) of a nucleosome and the total (writable) memory size of chromatin. I refer to Appendix A.1 for details what the IC represents and how it is calculated.
First, Bryant [13] provided a back-of-the-envelope calculation for the lower bound on the size of the human chromatin computer of individual nucleosomes and the full genome. This calculation was solely based on the approximative number of histone residues that can be modified and estimated to be 32 for the four core histones and 64 for the full nucleosome (because each of the four core histones is present twice). Bryant also deliberately ignored the existence of multiple distinct histone PTMs at particular residues. In her calculation, each nucleosome then contributes 64 bits

of information. The number of nucleosomes was estimated as 10 million, yielding $\approx$ 80 MB of memory.
Second, Prohaska et al. [12] also provided an estimate for the IC of individual nucleosomes but not for the full genome. Their calculation was based on a comprehensive list of reported mammalian histone PTMs that was collected from the literature. They calculated the IC of nucleosomes as $\approx$ 200 bits, which according to the authors constitutes up to one third of the total information stored on a chromosome (in other words, DNA makes up two thirds of the total information and epigenetic information one third).

However, both calculations are subject to various limitations. First, neither Bryant [13] nor Prohaska et al. [12] included the IC of DNA methylation and histone H1 although they both contribute to the memory capacity of chromatin. Second, both calculations are outdated because they, for example, do not incorporate newly discovered types of histone PTMs such as various lysine PTMs (succinylation [117], malonylation [117], crotonylation [116]) and tyrosine hydroxylation [116] as well as numerous additional previously undescribed histone PTMs as identified by Tan et al. [116]. Until October 2013, at least 223 distinct histone PTMs have been described for human alone (191 for the four canonical histones H2A, H2B, H3, and H4 as well as 32 H1.2 PTMs, see Appendix A.1 for details). Third, the calculation of Bryant is overly simplistic, whereas the estimate of Prohaska et al. is based on all known mammalian PTMs (i.e., not species-specific) and therefore difficult to compare.

Because researchers continue to identify more and more histone PTMs, the real memory capacity of chromatin was inevitably underestimated by previous calculations. In what follows, I provide an updated estimate of the IC of individual histones, nucleosomes and the total writable memory size of the full genome. I also provide an estimation of the theoretical upper limit, which has, to the best of my knowledge, not yet been done before.

In Table 3.2, I summarize the results of the memory analysis. The full details how I calculated these numbers are provided in Appendix A.1. The results suggest that the estimated theoretical memory size for the CC lies in the realms of several hundred megabytes of writable information per cell (Table 3.2). From the core histones that make up the nucleosome, H2B has the highest theoretical IC, due to the comparatively large number of lysine residues (which have the highest IC, see Table 3.1). However, based on the list of 223 known human histone PTMs (see above and Appendix A.1), H3 has the highest IC, partly due to the comparably large number of lysine residues that may be methylated. The information not encoded in DNA itself is non-negligible and estimated as 42-82% and 32-69% for individual nucleosomes and the full genome, respectively. This difference arises due to the presence of nucleosome-free regions. The IC for the CC is also much higher than for TF networks. Prohaska et al. [12], for example, estimated an upper bound for the IC of TF networks and argued that they have no more than 10^5 bit of information (i.e., $\approx$ 0.001 Mb). This calculation was based on an approximative number of regulators (half of all

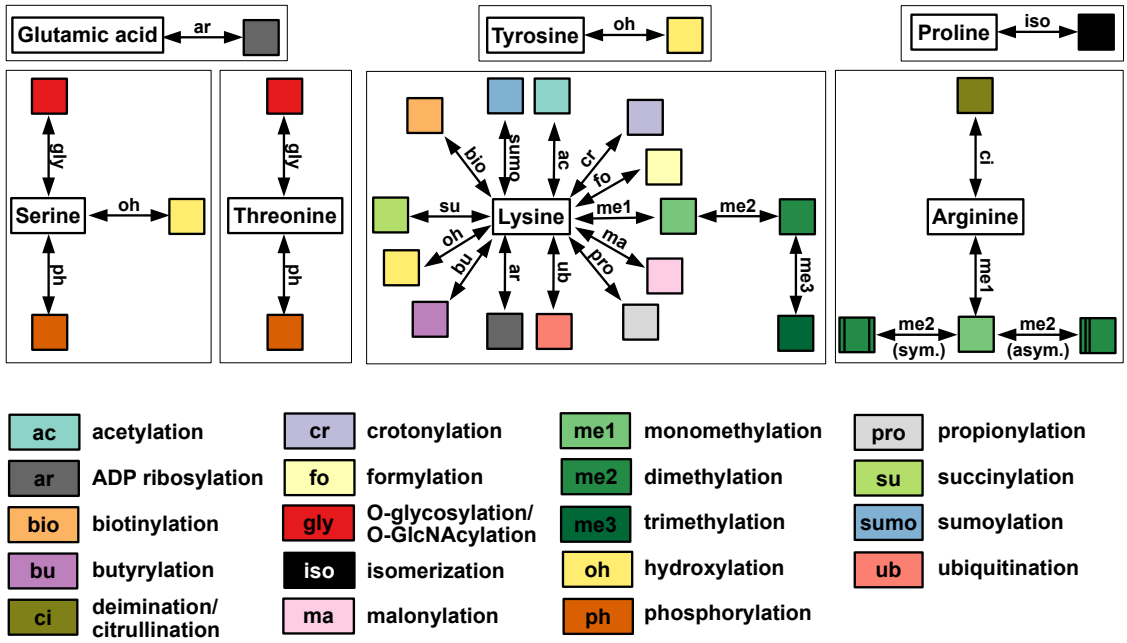


Figure 3.5: *Overview of all known types of histone PTMs and the amino acids where they occur. As of September 2013, researchers identified only seven amino acids to carry a histone PTM, and lysine is by far the most flexible residue (15 known histone PTMs), followed by arginine (4) and serine (3). For clarity, each type of histone PTM is colored distinctively. Note that as stated in the text, histone PTMs seem to be generally reversible.*

human genes; i.e., 10,000) and 1,000 copies for each regulator[6]. Even if one assumes 100,000 different regulators (counting alternatively spliced isoforms separately), each of which contributes 100,000 states (individual TFs may indeed be present in very high concentrations [289]), the total memory size of TF networks is still only ≈ 0.2 Mb. In reality, however, because TF networks display ultrasensitivity or threshold behavior (see Section 2.3.3), not each molecule represents a distinct state and consequently, the number of states for each TF can likely be encoded with only a few bits. One bit may even be sufficient because it is either absent or present in only such low concentrations that it has a negligible effect (state 0) or present with a concentration above a specific threshold so that the effect is non-negligible (state 1).

3.2.4 Computational Power and Efficiency

In this section, I analyze the computational power and efficiency of the CC. For this purpose, I compare the CC to the well-known TM, one of the standard models in computability theory (see Section 3.1.1). For this purpose, I will first formally define both a TM and a CC. For a formal

[6] 10^4 regulators and 10 bits per regulator to encode its concentration, yielding $10^4 \times 10 = 10^5$ if one assumes that each copy equals a state that is distinguishable by the cell

Amino acid	# distinct states	IC (in bits)	Known histone PTMs
Lysine	16	4	acetylation; mono-, di-, and trimethylation; formylation; crotonylation; ubiquitylation; sumoylation; biotinylation; succinylation; malonylation; propionylation; butyrylation; 5-hydroxylation; ADP ribosylation
Arginine	5	2.3	mono-methylation; symmetric and asymmetric dimethylation; deimination/citrullination
Serine	4	2	phosphorylation; O-glycosylation/O-GlcNAcylation; hydroxylation
Threonine	3	1.6	phosphorylation; O-glycosylation/O-GlcNAcylation
Glutamic acid	2	1	ADP ribosylation
Proline	2	1	isomerization
Tyrosine	2	1	hydroxylation
All others	1	0	no known histone PTMs

Table 3.1: *IC for each of the 20 amino acids with regard to known histone PTMs. As of October 2013, only seven amino acids may be post-translationally modified. Lysine is by far the most flexible residue (15 known distinct histone PTMs), followed by Arginine (4) and Serine (3). The IC of a particular amino acid is calculated as the logarithm to base 2 of the number of (known) distinct states. Note that the absence of any histone PTM also counts as a state (see column "# distinct states") and that both lysine and arginine methylation each contribute three distinct states.*

definition and mode of action of a TM, see Appendix A.2.

3.2.4.1 Formal Definition of a Chromatin Computer

For simplicity and comparison, I use the same notation as Bryant [13]. She defined a CC as the following 3-tuple $< M, B, R >$:

1. M: Set of m possible chromatin marks
2. B: Blank symbol that represents the absence of any chromatin mark, $B \subset M$
3. R: Transition function $R = M_*^{nk} \rightarrow M_-^{nk}$, with $M_* = M \cup \{*\}$ and $M_- = M \cup \{-\}$

Similar to the definition of a TM (see Appendix A.2), M is finite and non-empty. R defines a set of read-write rules, each of which reads the marks of n adjacent nucleosomes and overwrites them with new ones according to the definition of the rule. Nucleosomes have a particular number of positions (here: k, which I hereafter refer to as k-chromatin), each of which contains a valid symbol from M and may be independently modified by the rule set. One may therefore consider these positions as individual histone PTMs. For simplification purposes of the writing of the rules, in the left side of the rules definition, the special symbol $*$ is used to read any mark $m \in M$, whereas in the right side of the rule, $-$ is used to indicate that the original symbol remained unchanged (only

Category	Lower limit	Upper limit
IC per histone (in bits)		
H1.2	26	332
H2A	24	123
H2B	32	157
H3	44	138
H4	30	101
IC per nucleosome (including H1, in bits)		
DNA	400	400
DNA methylation	0	460
Core histones	260	1037
H1	26	332
All	686	2228
All but DNA (%)	286 (42)	1828 (82)
Total memory size of chromatin (in Mb)		
DNA	738	738
DNA methylation	6	27
Core histones	309	1236
H1	31	395
All	1084	2397
All but DNA (%)	346 (32)	1659 (69)

Table 3.2: *Memory capacity of chromatin. The table estimates the IC for individual histones, nucleosomes, and the full genome for various categories (DNA, DNA methylation, core histones, and histone H1), based on two different calculations (lower and upper limit, respectively). Due to the complexity of the methods and the restricted space in this table caption, see Appendix A.1 for the full details how I calculated these numbers.*

used for positions that are not uniquely determined). Collectively, this defines the configuration of the CC at any time point in the computation.

A CC that operates on k-chromatin, with rules depending on n adjacent nucleosomes and reading and writing marks from a set of m possible marks, is in the following referred to as a (m, k, n)-CC (Figure 3.6). In the simulations in Chapter 4, for example, I used a $(3, 1, 2)$-CC with $M = \{0, 1, 2\}$ and $B = 0$.

Mode of operation of the CC is similar to a TM (see Appendix A.2). A finite number of nucleosomes may contain valid input symbols $m \in M$, whereas the remaining ones are blank (B). At each step, a particular number of rules may match to the current chromatin configuration at various nucleosomes, and one rule is selected randomly. The states of the corresponding nucleosomes are then updated. The CC therefore typically operates in a non-deterministic fashion. It halts if no rule

matches.

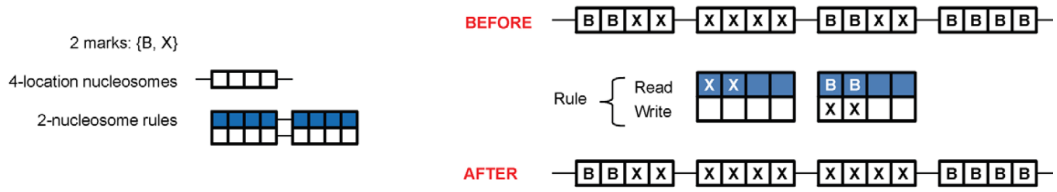


Figure 3.6: *Execution of a chromatin computer rule as defined by Bryant [13], which demonstrated the application of the* $(2, 4, 2)$*-CC rule* $[XX**][BB**] \rightarrow [----][XX--]$*. For clarity, each nucleosome is enclosed by* $[\ldots]$*. Figure taken from Bryant [13].*

3.2.4.2 Computational Power

In computability theory and computational complexity theory, analyzing the computational power of a particular model and proofs of computability and upper bounds on their computational complexity is typically performed by comparing it to well-known abstract models of computation such as the TM, cellular automata, or term rewrite systems. Most models of computations are Turing-complete (also called computationally universal) and therefore as powerful as TMs in terms of what they can compute. Examples include cellular automata [327], term rewrite systems (reviewed in [314]), and even computing with DNA [328] and membranes [302]. Although conceptually much simpler, string rewriting systems are also Turing-complete.

In addition, Bryant [13] recently showed that a CC is also Turing-complete and hence computationally universal. Briefly, she constructed a reversible mapping from any TM to 3-chromatin and from each Turing configuration to a chromatin configuration by transforming each cell from the tape of the TM to a nucleosome. In the mapping, each nucleosome has three positions (therefore 3-chromatin) to store (i) the current position of the head of the TM, (ii) its state, and (iii) the content of the current TM cell. Because the head of the TM is always at only one particular location, the first position of each nucleosome is empty for all nucleosomes but one. The same applies to the current state of the TM, which is also stored at the nucleosome where the head is located. Movements of the TM head are easily accomplished in the CC by moving the information of the current position of the head of the TM and its state to adjacent nucleosomes.

Bryant then showed that executing the CC results in a configuration of the CC that maps back to the Turing configuration that would have been achieved by executing the TM. Lastly, she showed that the CC halts whenever the corresponding TM halts. As an example for the computational power, Bryant solved the Hamiltonian path problem (Figure 3.1) with her chromatin model using three different algorithms, each of which has different requirements for memory space and computation time. For more details, I refer to Bryant [13].

In Chapter 4, I present a computational chromatin model that stays close to a physical model of chromatin. Due to its chemical realism, it contains a number of components that are not contained in the model of Bryant [13]. Nevertheless, I now demonstrate that the CC as presented in this thesis is also computationally universal. My strategy will be to reduce the presented CC model to the model of Bryant who previously proved Turing-completeness. For this purpose, I first compile a list of differences between the two models[7]. I then demonstrate that each of these differences is either a generalization of a concept already present in the model of Bryant or that it can also be integrated into the reversible mapping that Bryant established.

1. DNA replication
In contrast to the model of Bryant, the CC model as presented in this thesis includes DNA replication. However, this feature is an extension that may or may not be used in particular instances. For the proof of computational universality, we therefore assume that no DNA replication events occur (which is also biologically plausible, for example for amitotic cells such as neurons).

2. Reactions are associated with an explicit reaction rate
The CC model as presented in this thesis principally allows arbitrary reaction rates. They, however, have no significance for the reversible mapping because reaction rates only influence the frequency of selection in the Gillespie algorithm and the time increment. The former can be ignored because in the mapping, only one enzyme/reaction is applicable at any time (determinism), whereas the latter is irrelevant for the mapping (execution time is disregarded).

3. Enzymes may have arbitrary concentrations
In the model as presented here, enzymes and the rules they correspond to may have arbitrary concentrations, which affects their frequency of selection in the Gillespie algorithm. However, similar to the argument of the reaction rates, this has no significance for the reversible mapping. Due to the deterministic mode of action of the CC that results from the mapping, at any time point, only one type of enzyme (i.e., rule) is applicable at the specific position where the head of the corresponding TM is positioned. Consequently, it will be selected in the Gillespie algorithm regardless of its specific concentration.

4. Explicit distinction between binding and dissociation reactions
I model the dynamics of the "rewriting reactions" in a way that follows the mass action kinetics of chemical reaction systems, distinguishing, for each enzyme and each chromatin position, their elementary reactions. Bryant does not make such a distinction.
For retaining the reversible mapping between any TM and the CC, it must be ensured that the binding and corresponding dissociation reaction are executed consecutively. This can easily be achieved by converting each transition rule from the TM into two distinct rules in the CC as presented in this thesis when mapping from a Turing configuration to a CC configuration. The first rule corresponds to the binding reaction, whereas the second one realizes the state change and the

[7] see Chapter 4 for the specific model that is referred to here

dissociation. To force their consecutive execution, the CC has to be extended from 3-chromatin to 4-chromatin in the mapping (see above) by adding a new position to each nucleosome that stores if an enzyme is currently binding. By design, it thus has a value of 1 only for the nucleosome where the head is located and therefore where the binding of the corresponding enzyme occurred and 0 for all other nucleosomes (Table A.2). The first rule changes the position of this "bound" bit of the matching nucleosome to 1, whereas the second rule subsequently changes its state and resets the bound bit to 0. The left side of the second rule is therefore identical to the left side of the first rule, except that it requires the new fourth position to be 1 instead of 0 (which was set to 1 by rule 1 in the step before). Thus, because the second rule is the only that matches, the consecutive execution is forced by design. For an example TM program and the corresponding solution for the CC that implements the explicit distinction between binding and dissociation reactions, see Figure A.1 and Table A.2.

I now provide a generic mapping from a TM transition to the corresponding set of rules in the CC. For this purpose, I use one of the standard notations for a TM and their transitions (see Appendix A.2 for details). TM transitions that move the head to the left,

$$\{q_i, \gamma_k\} \rightarrow \{q_j, \gamma_l, L\} \text{ with } q_i, q_j \in Q \text{ and } \gamma_k, \gamma_l \in \Gamma,$$

convert to the following two rules in the CC:

$$1 : [\mathtt{BB} * 0][\mathtt{H}q_i\gamma_k 0][\mathtt{BB} * 0] \longrightarrow [\mathtt{BB} - 0][\mathtt{H}q_i\gamma_k 1][\mathtt{BB} - 0]$$
$$2 : [\mathtt{BB} * 0][\mathtt{H}q_i\gamma_k 1][\mathtt{BB} * 0] \longrightarrow [\mathtt{H}q_j - 0][\mathtt{BB}\gamma_l 0][\mathtt{BB} - 0]$$

TM transitions that move the head to the right are constructed analogously. Finally, TM transitions that do not move the head (head stays at position),

$$\{q_i, \gamma_k\} \rightarrow \{q_j, \gamma_l, S\} \text{ with } q_i, q_j \in Q \text{ and } \gamma_k, \gamma_l \in \Gamma,$$

convert to the following two rules in the CC:

$$1 : [\mathtt{BB} * 0][\mathtt{H}q_i\gamma_k 0][\mathtt{BB} * 0] \longrightarrow [\mathtt{BB} - 0][\mathtt{H}q_i\gamma_k 1][\mathtt{BB} - 0]$$
$$2 : [\mathtt{BB} * 0][\mathtt{H}q_i\gamma_k 1][\mathtt{BB} * 0] \longrightarrow [\mathtt{BB} - 0][\mathtt{H}q_j\gamma_l 0][\mathtt{BB} - 0]$$

In summary, the CC model as presented in this thesis may therefore be regarded as an extension of the model of Bryant [13] that retains the property of computational universality.

3.2.4.3 Computational Efficiency

As compared to traditional TMs, the CC achieves computational efficiency by various "extensions":

- Non-determinism: Different transitions may be applicable at a particular position (e.g., chromatin-modifying enzymes may acetylate a particular histone at different positions and they usually have multiple target nucleosomes to bind; vice versa, a particular nucleosome may be bound by multiple identical or heterogeneous chromatin-modifying enzymes)
- Probabilism: Choosing between multiple available transitions occurs according to some probability distribution (e.g., according to their free energy)
- Stay-option: The head may stay in the same position (e.g., chromatin-modifying enzymes may remain bound after a particular reaction)
- Multiple tapes: Each chromosome represents an individual and independent tape
- Multi-track: Each position in the tape consists of multiple symbols that correspond to the different amino acid residues that may be post-translationally modified
- Multi-head: One tape is associated with many heads that work independently and in parallel (e.g., multiple copies of particular chromatin-modifying enzymes)
- Random access: The head may not necessarily process the tape sequentially, it can also randomly move to any position on the tape (e.g., induced by higher-order structures, which bring distant nucleosomes in close proximity)
- Storage: Each transition is associated with a storage (e.g., the chromatin-modifying enzyme itself may be modified in its structure, composition, and modification state, all of which represents information that can be interpreted by the cell)
- Neighbor-dependency: Transitions not only depend on the state of the cell the heads points to but also on an arbitrarily defined "neighborhood". Different transitions (i.e., rules) may also require different sizes of the neighborhood (Figure 3.4)

3.2.5 Comparison of Ordinary, Chromatin, and DNA Computers

Various authors compared the similarities and differences between ordinary computers and naturally occurring biological computers in great detail (reviewed in [7, 9, 10]). Although these comparisons are out of the scope for this chapter because they generally apply to any biological computer and not just the CC, I want to highlight a few particularly noteworthy similarities and differences. First, the CC is an excellent example of a collection of components that were assembled during evolution on an as-need basis with a dynamic and changeable architecture, as compared to regular computers which have a rigid and fixed architecture. Consequently, the CC evolved in a modular

structure, which is also a key design principle for most programming languages as well as in complex networks (reviewed in [10]). Similar to decentralized and distributed network architectures such as peer-to-peer, the CC has no central control. Lastly, the CC flexibly operates through a large number of largely independently working processors, which is in difference to ordinary computers where fixed communication architecture severely limits both the variety of modes of operation as well as the number of processors that can be coordinated and programmed in an effective and fault-tolerant fashion [7].

Although DNA computers can solve specific problems[8], solutions were so far often time-consuming, laboratory-intensive, closely tailored to the problem, and rarely use a deterministic exploration of the search space [13]. CRMs, however, are much easier to implement but they are often limited to a subset of problems and typically cannot be programmed to solve arbitrarily complex problems [13]. The theoretical work from Bryant [13] suggests that the CC has the potential to be better suited to solve general-purpose programs (e.g., the Hamiltonian path problem) while also being computationally universal.

In terms of memory capacity, both a DNA and a chromatin computer are not restricted to a planar layer like general digital data [330] and are very promising candidates for high density storage of information. Furthermore, information storage in DNA is extremely stable. This is particularly true if considered double-stranded due to the increased stability and better readout [330] although it divides the theoretical capacity in half. The existence of various DNA repair mechanisms furthermore increases stability. For chromatin, it is more challenging to make efficient use of the theoretical memory capacity due to various reasons. First, the highly complex structure of histones and the interdependencies among individual histone PTMs and their corresponding reader, writer, and eraser molecules makes it more difficult to specifically write and retrieve the modification state of individual histone PTMs. Second, most histone PTMs are not particularly stable and have a short half-time (see Section 2.1.4.1), which poses a problem for long-term and maybe even short-term storage if not carefully designed. DNA methylation generally is more stable but may also be removed via various passive or active mechanisms (see Section 2.1.4.6). Third, information on chromatin is stored with high levels of redundancy to counteract the stochastic and highly dynamic mode of action, therefore decreasing the effective and usable memory size. Lastly, it seems unlikely that similar high accuracy repair mechanisms exist for chromatin.

3.3 Discussion

In this chapter, I analyzed composition, properties, mode of action, and computational power of the CC in a thorough and systematic fashion. I showed that the CC is Turing complete and therefore, at least in theory, able to not only perform computations in a biological context but also in a strict

[8] e.g., see [329] who implemented a particular matrix multiplication algorithm

theoretical informatics sense. As opposed to the traditional TM with an infinite tape, the CC has only a finite memory capacity. However, this similarly applies to all other naturally occurring computational models such as computing with DNA and membranes as well ordinary silicon-based computers that have been shown to be Turing complete. This limitation is, however, merely a theoretical one and does, in principal, not limit the computational power because any decidable problem that may be solved by a TM only requires a finite amount of memory. Thus, as long as the memory that is required for the problem is sufficient, no limitation exists.

Any problem that can be computed by a deterministic or non-deterministic TM may therefore also be computed by the CC with a specifically constructed rule set. Deterministic TMs are effectively a special case of non-deterministic ones, and due to their equivalency in terms of what can be computed, each non-deterministic TM can be simulated with a deterministic one. Different ways exist how this simulation may be performed, for example by using a 3-tape deterministic TM. Because the CC naturally contains multiple tapes (each tape abstractly correspond to one particular histone PTM), such a mapping is relatively straightforward to construct. Alternatively, for particular combinatorial problems that include a random selection of some kind, a CC may be constructed that resets the configuration whenever it recognizes that the proposed solution is not valid, thereby guaranteeing that the real solution is found in a single run in finite time (if a solution exists). This strategy has been implemented by one of the three CC solutions to the Hamiltonian path problem in Bryant [13]. The difficulty may be to formalize when a solution is valid, which will hardly be possible for all problems.

In the simulations as presented in Chapter 4, I interpreted all of the rules as symmetric with respect to their left and right neighbors as it seems not possible for the enzyme complexes to determine directionality (e.g. towards the centromere) from the local chromatin structure. In the proof of Turing completeness in this chapter, however, I assumed that rules are asymmetric. In reality, it appears that asymmetries are induced locally and thus are oriented relative to features such as promoters, insulators, and transcription factor binding sites. More generally, directionality of chromatin-modifying enzymes and their corresponding rules may be coupled and dictated by transcriptional directionality, which itself is influenced by a number of factors such as gene loops [331] and by ATP-dependent chromatin remodeling [332]. For example, directionality may simply be enforced by reposition or removing nucleosomes left or right to the position where the chromatin-modifying enzymes may perform his designated reaction. However, the proof of Turing completeness may principally also be performed with symmetric rules by adding a new position to each nucleosome that stores information about directionality.

I drew special attention to the execution unit that provides the logical and arithmetical operations of the CC and memory. Both are essential for the generation of complex, state-dependent responses [304]. Logical operations may be represented as read-write rules that are executed by myriads of "processors" — chromatin-modifying enzymes (see Chapter 4). As already highlighted in Chapter 4,

read-write rules may either depend or not depend on the modification state of other histone residues on the same or neighboring nucleosomes (context-independent and context-dependent rewriting rules, respectively). Due to the potential presence of multiple reader domains, the semantic complexity of these rules can be arbitrarily high although little is known about the true biological complexity for large multisubunit complexes. However, it is established that they can implement elementary logical operations that are analogous to `AND`, `OR`, or `XOR` gates in digital circuits. Chromatin therefore provides a potent and universal "language" in which computer programs or biological procedures may be written [13].

lncRNAs are also important for the logical operations of the CC by guiding chromatin-modifying complexes to particular genomic loci and therefore contributing to the regulation of gene expression (see Section 2.1.4.3). Rather than representing simple scaffolds, they may even represent complex "computer circuit boards" that link together various molecular components [316]. A CC therefore has its own logic, independent of the logic from pure CRMs.

The computational abilities of each chromatin-modifying enzyme seem to be very limited. They also have no a priori knowledge of their specific location although they may be anchored by lncRNAs or DNA binding proteins that are part of their complex. Furthermore, they work asynchronously and in a massively parallel fashion. Collectively, the similarities to amorphous computing systems are striking. Indeed, amorphous computing systems are widespread in biology and occur, for example, in neuronal networks, evolution, and organism development (e.g., morphogenesis) [11, 324]. Chromatin adds yet another example of a naturally occurring amorphous computation system that has not yet been described previously in that context.

I showed that a eukaryotic CC provides an enormous memory capacity that is potentially larger than the genome, which has also been noticed by others [12, 326]. Memory is predominantly realized by PTMs of histones and particularly histone tails and their specific recognition by proteins, and the sheer number of histone PTMs that are already known to occur indicate that the cell indeed makes use of it. Thus, as also noticed by Walker et al. [333], DNA only encodes a fraction of the total information in a cell. Information is not only stored in the DNA but instead in the current state of the entire system that particularly includes chromatin and its various forms of memory cells. Information not stored in DNA may thus serve as an extension of the IC by which the underlying genome is interpreted [334]. However, as noted in previous chapters, information on chromatin is stored with high levels of redundancy to counteract the stochastic and highly dynamic mode of action, therefore decreasing the effective and usable memory size.

The existence of histone variants (see Section 2.1.4.5) does typically not change the IC of a nucleosome because they replace the canonical histones with proteins of often very similar lengths and amino acid distributions. In reality, however, their incorporation often result in major structural changes of the nucleosome and therefore differential recognition by particular reader molecules. Furthermore, various dynamic chromatin remodeling phenomena may (temporarily) decrease the

memory capacity of chromatin such as histone tail clipping, the binding of proteins or ncRNAs to particular residues, histone exchange, and nucleosome eviction (Figure 3.3 C).
The IC estimation may, however, have only limited explanatory value. Cells are semantic systems[9] [335], and epigenetic information may not be stored in an information-theoretic sense [326]. Furthermore, the estimated theoretical upper limit is in practice smaller due to multiple reasons. First, some amino acid residues do not carry a particular histone PTM because it may induce an unfavorable change in the three-dimensional structure of the protein. Second, histone PTMs located in the histone fold are generally rare due to the crucial importance of these residues to maintain a functional structure. Third, particular combinations of histone PTMs are mutually exclusive and sterically impossible due to their spatial proximity.
In terms of utilizing the memory capacity for synthetic engineered systems, DNA already provides a stunning amount of useable memory. For example, Church et al. [330] recently encoded a 5.27 megabits bitstream (i.e., $\approx$0.7 Mb of information) onto DNA[10], which converts to a theoretical 5.5 petabits or around 700 terabytes per mm^3. As shown, chromatin has the potential to even increase that limit.

Memory in the CC may dynamically change the structure, content and properties of individual memory cells and is therefore less stable than DNA. To minimize the possibility of failure or misregulation and to accommodate for such high versatility, the cell employs much higher levels of redundancy for chromatin than for DNA. This is also in stark contrast to the mode of action of ordinary computers, which are constructed in a more streamlined design [336]. As for any chemical and biochemical information processing, noise control is of fundamental importance, and robustness is ensured by backing up regulatory states multiple times [7]. In the decentralized chromatin system, this is implemented, for example, through functionally redundant histone PTMs, particular higher-order chromatin structures, or altered transcription of genes coding for chromatin-modifying enzymes. Additionally, histone PTMs are arranged hierarchically to facilitate cellular modulation (i.e., primary histone PTMs are established first, whereas more secondary histone PTMs depend on the primary ones and are subsequently set) [336]. They therefore establish a dynamical and autonomous system of regulatory cascades that is "superimposed onto and uninfluenced by the underlying DNA" [12, p. 14]. However, various cellular components, such as ncRNAs (see Section 2.1.4.3), have the capability to anchor these chromatin-associated signals to the underlying DNA [12].

From an evolutionary perspective, redundancy may have naturally emerged from duplication and divergence [10] and from the observation that different components of the CC have strikingly different ages and origins. This is best exemplified by histones and particularly histone tails, the latter of which are only present in *Eukarya*. Composition and mode of action of the CC are therefore inevitably a product of evolution.
The chromatin regulatory system must allow that cells can alter their fate to produce distinct

[9] i.e., molecular information is associated with a particular meaning
[10] for details how these theoretical numbers were calculated, see Church et al. [330]

cell types during differentiation (plasticity). Additionally, cells must maintain this state when further differentiation is not necessary (inheritability) [139]. Inheritability must be achieved despite the fundamentally dynamic nature of the chromatin regulatory system in the form of remodeling processes and frequent DNA replications, the latter of which partly resets the stored information whenever the cell replicates. Inheritability and plasticity, however, are difficult to reconcile with one another [139]. The maintenance of correct epigenetic patterns throughout the lifetime of an organism is crucial for cellular stability and identity, and any misregulation of epigenetic mechanisms is likely to have fatal consequences. Indeed, it has long been speculated that an erroneously working CC is a common source of pathology that significantly or even etiologically contribute to diseases such as cancer and AD (see Chapter 6).

Chapter 4

Epigenetic Inheritance as a Computational Pattern Reconstruction Problem

4.1 Motivation and Background

In previous chapters, we have seen that the coupling of reading and writing of histone PTMs, which is a common feature of many histone-modifying proteins especially in crown-group eukaryotes, may have converted chromatin into a powerful computational device capable of storing and processing large amounts of information [12]. I also highlighted a recent theoretical study that showed that a simple model of chromatin computation, very similar to that proposed in Prohaska et al. [12], is computationally universal and hence conceptually more powerful than the logic circuits of cis-regulatory networks [13]. Although it is plausible that the computational capacities of chromatin play a role in the integration of external environmental signals and internal status information and hence in cell-fate decision, these computational aspects have remained largely unexplored. It is thus still unclear to what extent the potential power of chromatin computation is harnessed in real biological systems.

We have seen that the inheritance or recomputation of histone PTMs is not a trivial achievement since replication is associated with the partial replacement of histones and the deposition of newly assembled and hence unmodified histones [259, 260]. In other words, replication and the subsequent reconstitution of chromatin constitutes a dramatic disruption of the chromatin states that amounts to a partial erasure of the information stored in histone PTMs.

Here, I propose that the reconstitution of local patterns of histone PTMs is one of the biologically important computational tasks that is naturally solved by the "chromatin computer" (CC). The need to propagate epigenetic information to subsequent generations comes in two variants. The more stringent version concerns stable heritable bistability in which epigenetic information can be transmitted for, in principle, an infinite number of generations, depending on the strength of the bistability of the underlying system. As shown by Dodd et al. [36], this requires cooperative, positive feedback recruitment reactions as well as non-local interactions. In that regard, Rohlf et al. [17] provided a review of various modeling approaches for the dynamics and propagation of histone

PTMs. In contrast, the maintenance of local patterns of histone PTMs over a limited number of somatic cell divisions can potentially violate the conditions for long-term stability and tolerate slow accumulation of errors. In this setting it therefore makes sense to dispense with the stringent requirements outlined by Dodd et al. [36]. Indeed, it appears that this less stringent version is the relevant mechanism in multi-cellular organisms because cells only replicate a limited number of times (the Hayflick limit). Furthermore, particular epigenetic modifications can be gradually changed over generations through a number of different settings. In particular, these include the progressive reduction of higher histone methylation levels to lower methylation forms [271], epigenetic reprogramming [337, 338], epigenetic silencing / heterochromatin formation [339, 340], and transcription-coupled histone modifications [96]. Furthermore, researchers observed histone PTM gradients for a number of different modifications along a particular genomic region (for a review, see [82]). This effect may also indirectly result from multiple cell division due to the preferential retention of parental histones at the 5' end of genes [260]. Gradual changes of histone PTM levels may play a crucial rule in aging [170, 341]. DNA methylation changes, finally, are also are intimately linked to histone modifications [339], and therefore may be a direct result of the dilution of one or more histone PTMs.

One of the best-studied mechanisms proposed for epigenetic memory is based on positive feedback loops in nucleosome modification [278, 279]. The coupled reading, writing, and erasing of histone PTMs is therefore of crucial importance. Given the multiple tasks histone PTMs are involved in, I hereafter focus on whether a simple CC is capable of solving the pattern completion problem for a diverse set of chromatin input states, despite the highly disruptive nature of frequent cell divisions. More to the point, I ask whether it is feasible to find combinations of reader/writer enzymes that are capable of propagating, with high accuracy, preset chromatin states across several cell divisions. To answer this question I implement a generic stochastic simulation of rule-based chromatin modifications as a model of the CC. I then employ an evolutionary algorithm (EA) to evolve "programs" representing mixtures of rewriting rules to solve various pattern reconstruction tasks.

4.2 Methods

4.2.1 A Coarse-Grained Chemical Model of Chromatin Computation

I begin by introducing the computational model of chromatin. Computation abstractly consists of a system of states and transitions between them. My intention is to stay close to a physical model of chromatin. Similar to much of the literature, I define a *chromatin state* as the set of chemical modifications of histone molecules (or their absence) located at specific genomic positions. Although this is a simplified view, the full biological complexity of the components that make up a particular

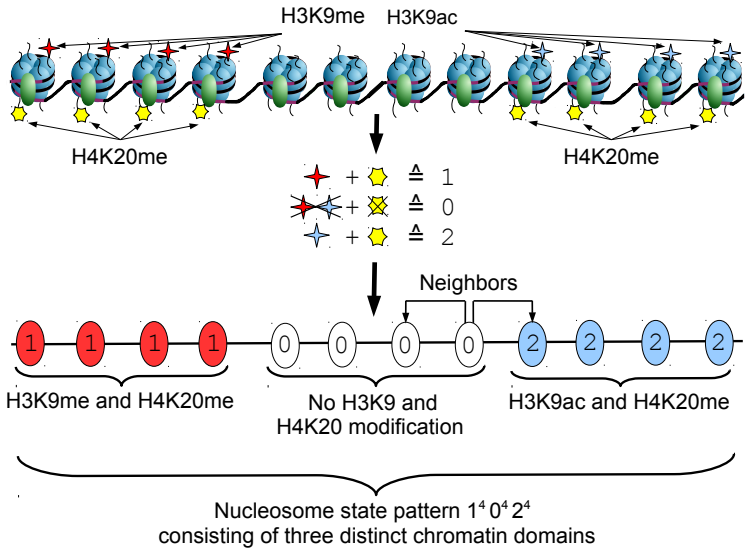


Figure 4.1: *Illustration of nucleosomes, their corresponding states and terminology that will be used throughout this chapter. A genomic region composed of 12 nucleosomes is shown, with H3K9 and H4K20 PTMs present at particular nucleosomes. For convenience, a translation table can then be used to assign combinations of histone PTMs (or the lack thereof) to particular symbols. Each nucleosome can then be assigned one of three distinct states (chromatin states): 0 (white), 1 (red), or 2 (blue). Homogeneously modified regions form a chromatin domain that carries a particular signature (0: unmodified H3K9 and H4K20, 1: H3K9me and H4K20me, 2: H3K9ac and H4K20me). Collectively, these 12 nucleosomes represent the local nucleosome state pattern or, abbreviated, simply pattern $1^4 0^4 2^4$. Such patterns therefore usually consist of multiple distinct chromatin domains (Table 4.3 for examples). The chromatin string uses a modified version of the Wikimedia Commons file "Nucleosome organization.png" (licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license).*

chromatin state cannot be integrated at this time given that the exact underlying mechanisms are incompletely understood. Therefore, I disregard effects that can have an impact on gene expression [342, 343] such as changes in nucleosome occupancy, the presence of histone variants, the effects of the three-dimensional structure of chromatin and nucleosomes (e.g., see [344]), chromatin remodeling events that increase the dynamic nature of chromatin (histone turnover, histone tail clipping, histone passback; see [175, 260]), and DNA methylation as an epigenetic mechanism that is known to at least partly interact with histone PTMs. I suggest that these details of the underlying "hardware" are not required for the investigation of the capabilities of the computational paradigm of the CC as a stochastic rewriting system.

These simplifications allow us to view chromatin as a linear sequence of n_n nucleosomes, analogous to the well-known "beads-on-a-string" picture of chromatin. Consequently, each nucleosome (except the two boundary nucleosomes) has two immediate adjacent neighboring nucleosomes or abbreviated simply "neighbors" (see also Figure 4.1). Because each nucleosome is completely specified by the collection of histone modifications that it carries, it can be represented by a single character $a \in \mathcal{A}$

that encodes its modification state (nucleosome or chromatin state). The symbol 0 is reserved for the unmodified state. As a technical simplification to save memory and to expedite the recognition of patterns in practical simulations, I use a single character instead of a string to represent the state of histone PTMs. I define a chromatin domain as a set of adjacent nucleosomes that are in the same modification state $a_i \in \mathcal{A}$. Each chromatin domain has a particular characteristic signature (e.g., methylated H4K20 and acetylated H3K9 residues) and length l and therefore can be represented by a sequence of nucleosome states a^l. I call this a local nucleosome state pattern (or simply pattern) to emphasize that the chromatin state is solely determined by the nucleosome state. Naturally, such patterns may also be composed of multiple adjacent distinct chromatin domains (e.g., see Figure 4.1 and Table 4.3).

Transitions between chromatin states are mediated by histone-modifying enzymes. These enzymes catalyze the writing or erasing of histone PTMs in a context-dependent manner. Therefore, they are implemented as string rewriting rules acting on the nucleosome state pattern. (Figure 4.2 A). They recognize parts of the (local) nucleosome state patterns and cause a change in the modification state of one or more nucleosomes. These rewriting rules are described in detail in the following section.

Not all histone-modifying enzymes are present in the cell at the same time or in same concentrations, and they may differ substantially in their affinity to their target patterns or in their catalytic efficiency. Furthermore, different enzymes may compete for the same chromatin locations and, *vice versa*, different chromatin positions may compete for low abundance enzymes (Figure 4.2 C). I therefore model the dynamics of the "rewriting reactions" in a way that follows the mass action kinetics of chemical reaction systems, distinguishing, for each enzyme and each chromatin position, their elementary reactions (Figure 4.2 B):

1. Binding of an enzyme to a specific locus (i.e., one or more nucleosomes). I assume that enzymes only bind to nucleosomes on the chromatin string that match the precondition of the rewriting rule that they embody. Also, enzymes cannot bind if any of the nucleosomes that are decisive for the applicability of the rewriting rule are bound by other enzymes, thereby blocking the accessibility.

2. Dissociation of an enzyme and the application of a rewriting rule (i.e., writing or erasing of one or multiple histone modifications). After dissociating, the enzyme is again available for new reactions.

Contrary to most previous work and simulation systems for chromatin state dynamics (e.g., see [36, 38, 39]), I employ a chemical reaction system and model each reaction explicitly. Additionally, the computational paradigm used here for the enzyme kinetics is a stochastic one. I argue that this level of chemical realism is crucial, since concentrations of regulatory molecules, rather than their mere presence or absence, are very well known to be of crucial importance in the regulation of gene

expression. Indeed, few regulatory events are qualitative — typically changes in expression levels of regulators are gradual and rarely exceeding a few-fold increases or decreases.

Chromatin state dynamics are thus dependent on enzyme abundances, the availability of local patterns on which they can act, the current state of the system, and the rate constants for each chemical reaction (Figure 4.2 C). The latter generally quantify the speed of a chemical reaction and may differ substantially among different enzymes.

Histone PTMs have strikingly different lifetimes and are deposited at different rates. Acetylation events are measured in the order of minutes, whereas methylation events are stable for days [173]. These rate differences are determined by the enzymes that catalyze the corresponding reactions [277]. A given mark can be removed either by specific de-modification enzymes or through chromatin remodeling (e.g., histone turnover or histone tail clipping). Since chromatin remodeling phenomena are at present not explicitly included in the model, different life-times can be modeled by neighbor-independent rewriting rules with different rate constants. To the best of my knowledge, spontaneous (i.e., enzyme-independent) decay has yet to be described for histone modifications although it cannot be excluded that some of the more exotic or yet undescribed modifications may not require an enzyme for de-modification.

For simplicity I use a single reaction rate parameter for the binding of histone-modifying enzymes or enzyme complexes although mechanistically, this may require multiple steps (e.g., binding, recruitment of other factors and oligomerization). The propensity for a particular binding reaction is computed as the product of its reaction rate and the number of free (i.e., not bound) molecules for that enzyme, whereas for any dissociation reaction the propensity equals its reaction rate and is therefore independent of the number of free molecules.

The time course of the simulation between replications can be subdivided into discrete "phases" (Figure 4.4) that can have varying durations. For each phase, enzymes can be arbitrarily set to be present or absent (hereafter denoted enzyme availability) (Figure 4.4). However, enzyme concentrations either have a constant value as specified by the programming of the CC (present) or a value of 0 (absent). If an enzyme is still bound at the transition from one phase to the next, it dissociates without performing a state change at the nucleosome it binds to. Collectively, these phases abstract the cell's gene expression program. I may interpret them for instance as the G1, S, and G2 phase of the cell cycle. Alternatively, phases may be used to model distinct developmental stages.

I also include replication in the model. At regular time intervals, a replication takes place and the parental histones are distributed between the two daughter strands. As discussed in Section 2.2.4 and Figure 2.7, different models have been proposed for how histones are segregated after DNA replication. For the purpose of the present study, I adopted the random model (Figure 4.2 D) because it has the best experimental support. I treat nucleosomes as indivisible units, which, however, is only

a simplification if multiple histone PTMs from different histones are modeled. I note, however, that the simulation environment can easily be extended to other replication models if the need arises. Again, I argue that this additional level of biological realism is irrelevant for the questions addressed here. I also assume that parental histones are redeposited at their prereplication locus, which is consistent with the finding that most parental histones in budding yeast are reincorporated in close vicinity, i.e. within 400 bp, of their original locus [260]. Lastly, analogous to the phase change transitions, enzymes that are still bound at the time of replication dissociate without performing any reaction.

4.2.2 Chromatin Enzymes as Rewriting Rules

In the simplest case, histone-modifying enzymes evaluate only the state of the nucleosome that they modify. The model of Dodd et al. [36], for example, considers three distinct states: unmodified (0), methylated (M), and acetylated (A). Each state can be interconverted by the catalytic actions of histone acetyltransferases (HATs), histone deacetylases (HDACs), histone methyltransferases (HMTs), and histone demethylases (HDMs). The corresponding set of rewriting rules is

$$\begin{array}{ll} \texttt{HAT:} & \texttt{0} \rightarrow \texttt{A} \\ \texttt{HDAC:} & \texttt{A} \rightarrow \texttt{0} \\ \texttt{HMT:} & \texttt{0} \rightarrow \texttt{M} \\ \texttt{HDM:} & \texttt{M} \rightarrow \texttt{0} \end{array} \tag{4.1}$$

As opposed to these simple, neighbor-independent rewriting rules, more complex ones, such as the ones considered in Sneppen et al. [39] and Bryant [13], also depend on the neighboring nucleosomes (neighbor-dependent rewriting rules). In my implementation, arbitrarily complex rewriting rules can be specified (see Table 4.1 and Figure 4.2 for a few examples). For convenience, I allow wildcards in the definition of the rewriting rule and the state of the neighboring nucleosomes can be incorporated as well. Rules are interpreted as symmetric with respect to their left and right neighbors because it is not possible for the enzyme complexes to determine directionality (e.g. towards the centromere) from the local chromatin structure. In reality, it appears that asymmetries are induced locally and thus are oriented relative to features such as promoters, insulators, and transcription factor binding sites [345].

Such complex rewriting rules indeed seem to be common for eukaryotic systems because histone-modifying enzymes are often part of large enzyme complexes with multiple protein domains, each of which having a particular function such as DNA sequence recognition and histone binding. Individual histone PTMs in the vicinity of the binding may substantially alter the binding affinity of the enzyme complex or its formation. This crosstalk is frequent within histones, among histones of the same nucleosome [145–148], and among histones of neighboring nucleosomes (for an overview, see [149]). Even for single domain proteins, the presence of multiple modifications may be required for binding [224].

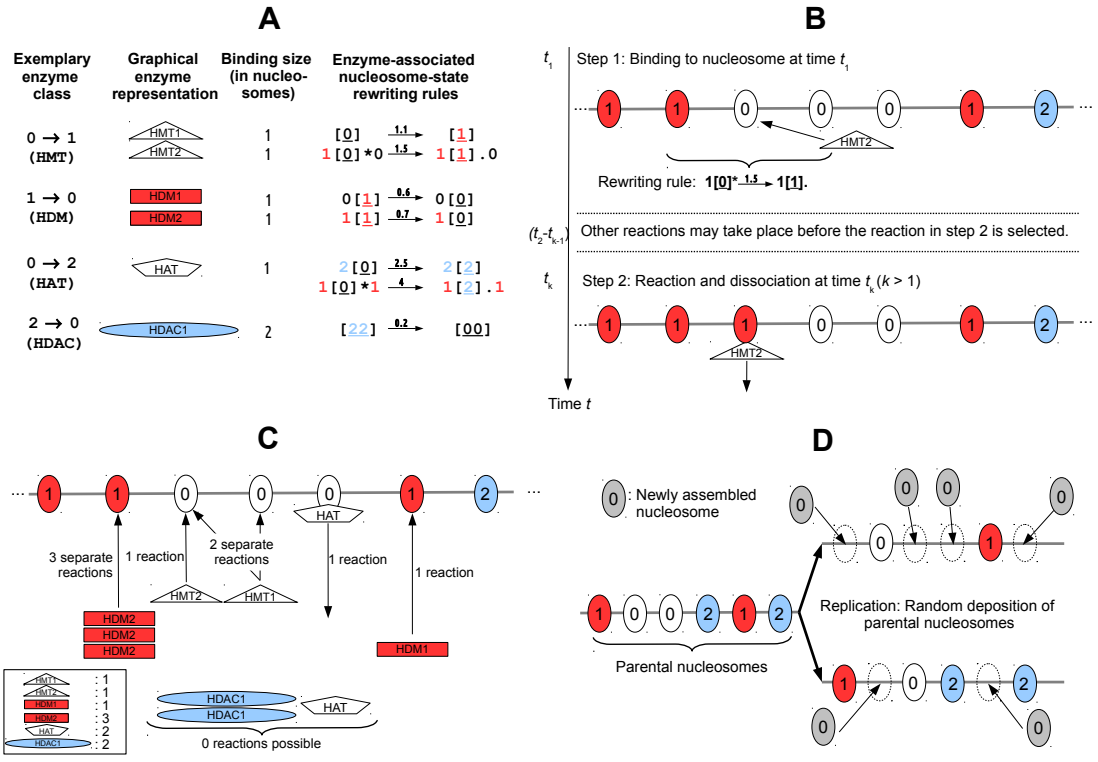


Figure 4.2: *Basic ingredients of the chromatin model. An example of a genomic region is shown where nucleosomes can either be in state 0 (white), 1 (red), or 2 (blue).*
A: Enzymes and rewriting rules. A total of seven enzymes are defined that can be broadly divided into the following four well-known classes (for illustration purposes): histone methyltransferases (HMTs), histone demethylases (HDMs), histone acetyltransferases (HATs), and histone deacetylases (HDACs). Each enzyme recognizes and binds to a particular, local nucleosome state pattern (embraced by square brackets in the rewriting rules) and thereby blocks accessibility of these nucleosomes for other enzymes. Such recognition patterns may be dependent or independent of neighboring nucleosome states and can be represented as rewriting rules. They may proceed at different rates and change the state of at least one nucleosome.
B: Enzymes and their reactions. Binding of an enzyme to one or more nucleosomes and the corresponding dissociation at a later time are modeled as separate reaction steps in the model. First, enzymes may bind to one or more nucleosomes as described by their rewriting rules. They then remain bound until the corresponding dissociation reaction is selected in the stochastic simulation. When the enzyme dissociates, the bound nucleosome(s) change their state(s), as specified by the corresponding rewriting rule(s).
C: Enzyme competition and reaction dynamics. All eight possible reactions that can occur for various enzymes and their corresponding concentrations are shown (see legend) as well as an exemplary genomic region with seven nucleosomes. Each arrow indicates a possible reaction that may take place at this particular time point. If enzymes are bound to particular positions (e.g., HAT), they block the accessibility of other enzymes at bound nucleosomes. Note that particular enzymes can be present multiple times (e.g., HDM2) and that they may be able to bind to the same nucleosome (e.g., HMT1 and HMT2), resulting in competition for nucleosome binding. Other enzymes may not be able to perform any reaction due to the inapplicability of their rewriting rules given the current state of the nucleosomes (e.g., HDAC1 and one of the two HAT molecules).
D: Replication and nucleosome segregation. The random distribution model of chromatin replication is shown (see Section 2.2.4 for details).

Table 4.1: *Example of different valid nucleosome state rewriting rules for a chromatin-based system with three possible nucleosome states $\mathcal{A} = \{0, 1, 2\}$. The symbol "$*$" in the left part of the rewriting rule is a wildcard matching any nucleosome state, whereas "." on the right part of a rewriting rule means that the nucleosome state is left unchanged. The nucleosomes bound and rewritten by the corresponding enzyme are embraced by square brackets.*

Rewriting rule	Matching patterns for the rewriting rule	Neighbor-dependent
$[1] \xrightarrow{r_1} [0]$	$[1] \xrightarrow{r_1} [0]$	No
$[(1\|2)] \xrightarrow{r_2} [0]$	$[1] \xrightarrow{r_2} [0]$, $[2] \xrightarrow{r_2} [0]$	No
$1[1] \xrightarrow{r_3} 1[0]$	$1[1] \xrightarrow{r_3} 1[0]$	Yes
$1[1] * 0 \xrightarrow{r_4} 1[0].0$	$1[1]00 \xrightarrow{r_4} 1[0]00$, $1[1]10 \xrightarrow{r_4} 1[0]10$, $1[1]20 \xrightarrow{r_4} 1[0]20$	Yes
$[11] \xrightarrow{r_5} [01]$	$[11] \xrightarrow{r_5} [01]$	No
$1[11]0 \xrightarrow{r_6} 1[10]0$	$1[11]0 \xrightarrow{r_6} 1[10]0$	Yes

I represent these complex rules as follows: Each rule is specified by a pair of strings, the left part is the template that must be matched to the local nucleosome state pattern. This precondition must be met so that the rewriting rule can be applied to the matched string. Once a match is found, the left string is replaced by the right string. I therefore regard the system of rules as string rewriting system (see Section 3.2.1.2).

In my implementation, both strings consist of up to three parts: (i) a string of length $l_1 \geq 0$ for the required state(s) of the neighboring nucleosome(s) left of the actual binding site of the corresponding enzyme; (ii) a string of length $l_2 > 0$ for the required state(s) of the nucleosome(s) that the enzyme binds to; (iii) a string of length $l_3 \geq 0$ for the required state(s) of the neighboring nucleosome(s) right of the binding site. Only the nucleosomes of the binding part may be subject to change in the right part of the rewriting rule (as indicated by a^* in the equation below). Thus, rules are not reversible per se; instead, reversible reactions must be formulated as separate forward and backward reaction. Furthermore, each rewriting rule is associated with a rate constant. In summary, it is formally specified in the following form:

$$\begin{aligned} a_1 \ldots a_{l_1}[a_{l_1+1} \ldots a_{l_1+l_2}]a_{l_1+l_2+1} \ldots a_{l_1+l_2+l_3} \xrightarrow{rate} \\ a_1 \ldots a_{l_1}[a^*_{l_1+1} \ldots a^*_{l_1+l_2}]a_{l_1+l_2+1} \ldots a_{l_1+l_2+l_3} \end{aligned} \tag{4.2}$$

where each $a_i \in \mathcal{A}$ denotes a particular nucleosome state. Table 4.4 compiles the rewriting rules with $l_1, l_2, l_3 \leq 1$ that are used throughout this chapter.

The nucleosome string can be either linear or circular. For the former, rewriting rules that require the state of both neighboring nucleosomes cannot match the two boundary nucleosomes. This entails that they keep their parental status unless replication replaces them with unmodified nucleosomes, after which their original state is lost permanently (Figure 4.7). In the latter case, the two boundary nucleosomes are directly connected, and rewriting rules may match.

4.2.3 Modeling the Dynamics of Histone Post-Translational Modification States

4.2.3.1 General Modeling Approaches for Chemical Reaction Systems

Researchers modeled a large array of systems in the past, ranging from protein folding to networks of metabolites to models of how the brain functions. The reconstitution of local patterns of histone PTMs by histone-modifying enzymes may also be regarded as a chemical reaction system. Systems can be modeled manifold, depending on the size of the system and associated time constraints in the simulation itself, the addressed questions, and the level of detail the model should include. Generally, system dynamics can be modeled using either deterministic or stochastic models. Before describing my approach, I present a brief overview of deterministic or stochastic models.

Deterministic approaches

Deterministic approaches generally always produce identical output for a given input. For chemical reaction systems, they typically refer to kinetic modeling of a set of ordinary differential equations (ODEs). Each ODE describes a number of reactions, time-dependent concentrations of the corresponding molecules are the variables of the system, and reaction rate constants represent the parameters. ODE approaches thus describe changes in amounts of components over time [289].

However, this approach neglects of dependencies on spatial locations (i.e., reactions occur homogeneously throughout the reaction volume) and assumes that the number of molecules is sufficiently large so that reaction discreteness and biological noise caused by stochastic processes have no macroscopic effect (i.e., reactions occur simultaneously). Thus, for biological systems, they usually do not constitute a physically accurate representation of the underlying processes. Therefore, researchers did not employ them for epigenetic inheritance models.

Stochastic approaches

One challenge of all biological systems is their stochastic nature. The simplifying assumptions of deterministic models break down if the number of molecules is small or if random fluctuations are non-negligible because reactions may then occur in a particular order rather than simultaneously and stochastic effects may have substantial influences on the system. Both of these phenomena may result in macroscopic effects that influences the future dynamics of the system [289, 346]. Thus, including stochasticity in the model is appropriate. This typically entails the replacement of differentiable concentrations of the ODE approach by state probabilities defined by the number of molecules of each type at a given time and that evolve in time. However, stochastic approaches generally require much more computation, which may not be always feasible. Nevertheless, due to the exponential increase in computing power, stochastic *in silico* modeling has emerged as a powerful means to improve understanding of complex systems.

Three common approaches for stochastic modeling of chemical reaction systems exist, and to the best of my knowledge, researchers only employed approximative CME approaches for epigenetic inheritance models (e.g., by using a mean-field theory ansatz [30–33] or transfer matrix ansatz [34]).

Chemical Master Equation (CME) approach. The CME precisely describes the time-evolution of a chemical reaction system. At any given time point, the system is in one particular state out of a countable number of states, with probabilistic switching between states. Analytically solving the CME for long time behavior is generally infeasible and often even mathematically intractable due to its complexity. For example, in contrast to the ODEs approach for which only one ODE per species is required, the CME approach requires one ODE per state of the system). Approximations, such as the mean-field theory, may, however, be used [30].

Stochastic simulation approach. Instead of dealing with the CME directly, one can make exact numerical calculations by using reaction probability density functions to determine when the next reaction will occur and what type of reaction will be executed. This approach is called the stochastic simulation approach. Unless approximative methods are used, it is rigorously equivalent to solving the CME of the corresponding stochastic model because it exactly represents the stochastic version of the trajectory of the corresponding CME that embodies the system. This approach is computationally very expensive because each reaction is explicitly simulated. Furthermore, because each simulation is subject to stochastic fluctuations, multiple independent simulations must be performed and analyzed collectively to fully understand the properties of the system.

Stochastic differential equations (SDEs). Lastly, stochasticity may be modeled with SDEs. SDEs generally combine ODEs with a stochastic component by mixing ODEs with random fluctuations such as Brownian motion. However, SDEs are difficult to analyze numerically, and solving SDEs is a relatively young field. To the best of my knowledge, researchers did not yet employ SDE approaches for epigenetic inheritance models.

4.2.3.2 Stochastic Simulation with the Gillespie Algorithm

As outlined in Section 2.3.2.2, the recruitment-copying model attracted particular attention. However, as pointed out in Chapter 1, existing approaches, such as the ones from Dodd and Sneppen [36, 38, 39], suffer from various limitations because they (i) utilize only very simplified mathematical descriptions for the system that do not explicitly model chemical reactions, (ii) do not incorporate cellular concentrations of chemical species, and (iii) use the simplifying assumption that the time evolution of the system is continuous. In addition, they use only a fixed set of reactions/rules that are suitable for the system *ab initio* and elementary histone PTM patterns that are ought to be retained. For example, investigating whether such a system can easily evolve a set of histone-modifying enzymes (i.e., rules) capable of stably inheriting a particular histone PTM pattern across cell divisions in the first place has so far never been addressed explicitly. Furusawa et al. [347] used

evolutionary simulation and optimization for a gene regulatory network that contained epigenetic feedback loops (precisely, feedback loops between gene expression levels and their epigenetic control) but this is entirely different from the model of epigenetic inheritance as discussed here. Their model therefore cannot answer the question whether epigenetic inheritance can be formulated as a computational problem.

For the chromatin system, I employ a stochastic simulation approach. Specifically, I use the stochastic simulation algorithm as proposed by Gillespie [348, 349] (also called first reaction method), which is a well-established, widely-used[1] and numerically exact algorithm for the stochastic modeling of cellular systems.

Requirements and Formalization

Conceptually, imagine a system with a finite number n_s of chemical species S_i ($i = 1, \ldots, n_s$), each of which is present a particular number of times. At each time point t during the simulation, this defines a vector $X(t)$ of dimension n_s representing the abundances of each chemical species at time t (population state vector). These n_s chemical species may interact through a finite number n_r of reaction channels R_j ($j = 1, \ldots, n_r$). Generally, each reaction channel R_j reflects any particular process that changes the population size of at least one chemical species in the system, and it is characterized by three quantities: its state-change vector (change of the concentrations of the available chemical species upon execution of R_j), its rate constant r_j, and its propensity function $a_j(X(t))$ (probability that one R_j reaction will occur in the time interval $[t; t + dt)$, see [349]). Such propensities are simply determined by multiplying r_j with the number k_j of chemical species that can execute the reaction. For example, for a typical binding reaction of a protein complex of type p to a particular genomic location l, k_j equals the number of available complexes of type p that can potentially bind l (as specified in $X(t)$). When the specific complex of type p later dissociates, k_j equals 1 because only this particular complex is involved in the reaction.

Algorithm

Given an initial time t_0 and population state $X(t_0)$, the algorithm proceeds as follows. In each iteration, two random numbers r_1 and r_2 are drawn from the uniform distribution in the unit interval. Whereas r_1 is used to determine the time increment τ, r_2 is used for the selection of a reaction channel. τ is calculated as the sum all reaction rates (propensities), i.e, $\tau = \frac{1}{a_0(X(t))} ln(1/r_1)$ where $a_0(x) = \sum a_j(X(t))$. If a particular reaction channel cannot be executed at a particular time, its propensity is 0 and it therefore cannot be selected. The index of the next reaction to be executed is the smallest integer j satisfying $j = \sum_{i=1}^{j} a_i(X(t)) > r_2 a_0(X(t))$. That is, a reaction channel is selected proportional to their respective weights (Figure 4.3). The reaction R_j is then executed, the state of the system changes according to the state change vector of R_j, and the current time t is

[1] as exemplified by over 2,500 citations in total and over 800 citations since 2010

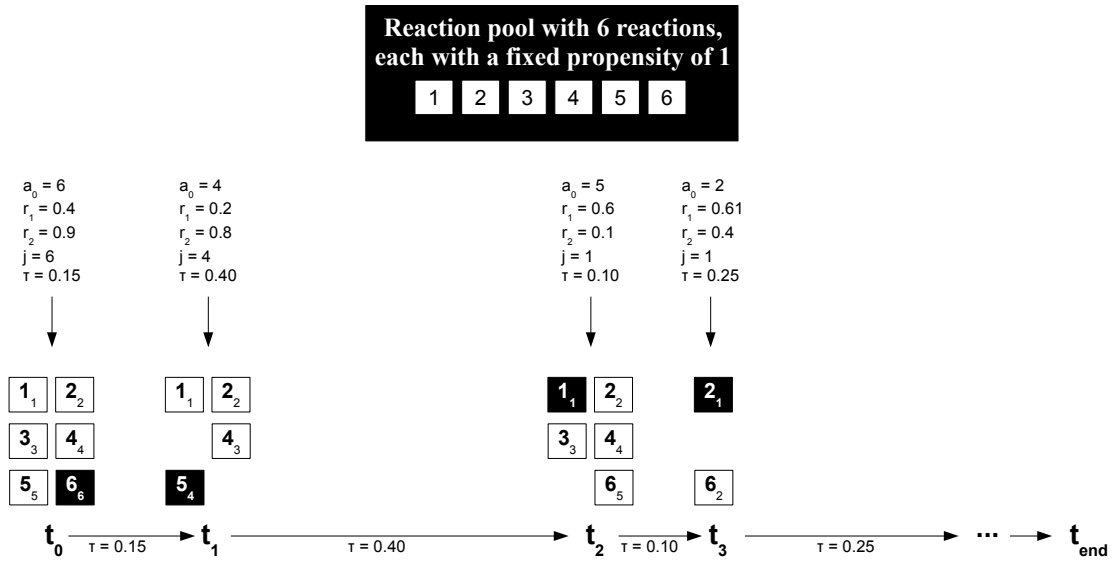


Figure 4.3: *Illustration of the stochastic simulation algorithm as proposed by Gillespie. An exemplary biological system with six possible reactions (numbered from 1–6, squares), each of which has a fixed propensity of 1 (for simplicity). The first four iterations of the stochastic simulation algorithm are shown, starting from time point t_0 to t_3. In each iteration, the available reactions, the values of the two random numbers r_1 and r_2, the sum of all propensities a_0 and the resulting values of the time increment τ and the reaction j that is executed next (highlighted in black) are shown. The numbers in subscript next to the reaction indexes illustrate the internal numbering of the reactions in that particular iteration (see the j value), which may differ from the original reaction numbers because some reactions may not be available in a particular iteration of the algorithm.*

subsequently incremented by τ. Thus, time increments are not fixed but instead of variable length, dependent on the current state of the system (i.e., the available reaction channels and population sizes of all chemical species). This process is repeated until a stop criterion is reached (e.g., a particular simulation time, either in terms of total computing time or time units in the simulation).

Limitations and Variants

The stochastic simulation algorithm also has limitations. First, although the algorithm is rejectionless and hence efficient in terms of the simulation, it is too slow for biological systems with large population sizes or the high reaction rates because the time increment τ between reactions decreases dramatically, thus slowing down the simulation substantially. Researchers proposed a number of computationally more efficient variants of the original stochastic simulation algorithm that improve the worst-case time complexity with respect to selecting the next reaction to execute from $\mathcal{O}(n_r)$ to $\mathcal{O}(\log n_r)$ [350] and even $\mathcal{O}(1)$ using a composition and rejection algorithm [351]. Gibson et al. [352] published an exact variant of the original algorithm (termed next reaction method) in 2000 that also scales in $\mathcal{O}(\log n_r)$ by reusing unused reaction times and making the sampling of reactions therefore more efficient. The variant by Slepoy et al. [351] is especially efficient because it also scales

in $\mathcal{O}(1)$ in the rest of the algorithm, and except for exceptionally large cellular systems, this calls the necessity to use approximative methods into question. Approximative methods, such as the explicit (e.g., see [353]), binomial (e.g., see [354, 355]), and optimized tau-leap method (e.g., see [356]), generally do not simulate one reaction at a time but instead a bunch of reactions. Various authors proposed additional approximative variants, such as stiff system methods, that improve efficiency for systems containing reactions with vastly different time scales [357] and adaptive methods for automatically choosing between explicit and implicit tau-selection methods for improved efficiency [358].

Another limitation occurs in cellular systems where chemical species with multiple states are used, as mentioned by Le Novere et al. [359] and analyzed more thoroughly by Liu et al. [360]. The stochastic simulation algorithm may not be applicable due to the sheer combinatorial explosion of all possible reactions that may take place. This is the case, for example, for proteins that can be post-translationally modified at multiple independent sites, each of which may influence the reaction rates of the complex [359]. If the number of possible states is high, the `StochSim` algorithm [359] may be an alternative to consider, despite its approximative nature [360].

Implementation

Since the set of available reaction channels and their weights are, in general, dependent on the current state of the system, the "book-keeping" of all reaction channels and their status is an important issue for the practical implementation of the Gillespie algorithm. This is particularly relevant for models with large numbers of different molecules and reactions. I next describe the peculiarities and design decisions that are specific to the chromatin-based model. To do so, I make use of abbreviations for relevant parameters that are summarized in Table 4.2. Consider a system with n_e enzymes, each of which has a particular number of rewriting rules. The sum of all rewriting rules is then $n_{\text{rules}} = \sum_{k=1}^{n_e} |e_k|$, where $|e_k|$ denotes the number of rewriting rules that are defined for that particular enzyme. In the present system, each rewriting rule adds a total of $2n_n$ reaction channels (one corresponding to the binding and one to the dissociation for each of the n_n nucleosomes). Thus, the total number of reaction channels is $2n_n \times n_{\text{rules}}$. Equivalently, each of the n_n nucleosomes has $2n_{\text{rules}}$ different reaction channels).

Internally, each rewriting rule is associated with a state change vector, which describes how the concentrations of the available n_l enzymes are affected upon execution of the reaction that specifies the rewriting rule. For example, any binding reaction decreases the concentration of the corresponding enzyme by 1, whereas any dissociation reaction frees the enzyme and thereby increases its concentration by 1.

To save computing time, only the propensities of the reactions in the vicinity of the nucleosome(s) subject to the last reaction are recomputed. A recomputation of the reaction propensities is also necessary after each replication and phase transition.

Each replication event occurs periodically after a fixed time interval t_r. Similarly, each phase p_i has a defined duration and stops after a particular time t_n. This requires a correction for the last reaction event in each period — i.e., the one for which $t + \tau > t_r$ and $t + \tau > t_n$, respectively. Here, I draw a random number r_3 and accept the reaction if $r_3 > (t_r - t)/\tau$ and $r_3 > (t_n - t)/\tau$, respectively.

4.2.4 Evolutionary Optimization of the Rewriting Rule Sets

The CC may operate with rather complex instructions that correspond to a particular gene expression pattern. More formally, an instruction consists of a list $\mathcal{L}$ of rewriting rules and associated enzyme concentrations (Figure 4.4). This is similar to the computational model of Bryant [13], except that I have incorporated a concentration associated with each rewriting rule that modulates the probability with which it is applied. If the time course of the simulation is divided into n_p phases, more complex programs can be implemented as sequences $(\mathcal{L}_i, \tau_i), i = 1, \ldots, n_p$ of instructions that are valid for a prescribed time period τ_i (i.e., a "phase") before being supplanted by the next instruction. Thus, if a particular rewriting rule $r_k \in \mathcal{L}_i$, the enzyme performing this reaction is available in phase i.

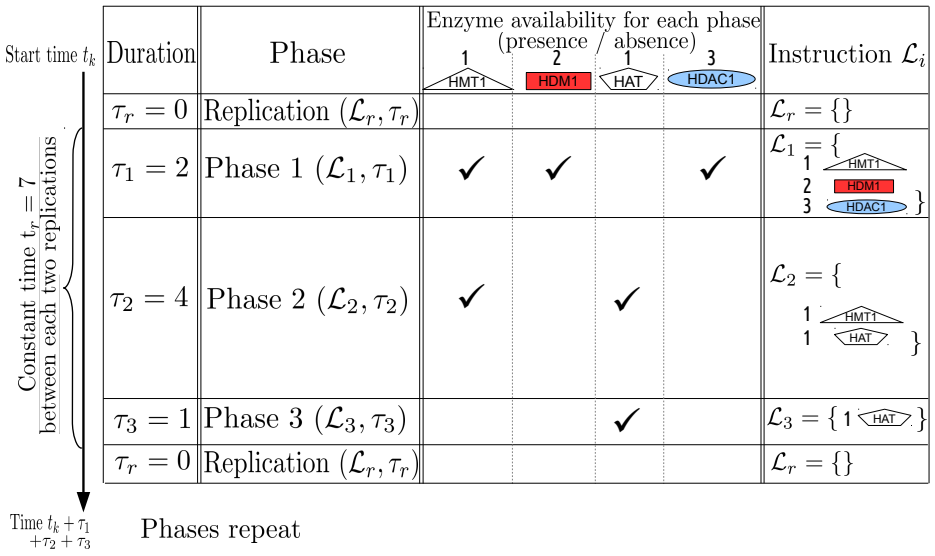


Figure 4.4: *Illustration of the phases concept. An example with four different enzymes (each of which has a particular concentration) and three phases is represented. After each replication, these phases periodically follow each other in sequential order. Formally, each phase i is described by a tuple $(\mathcal{L}_i, \tau_i)$. Each phase i can be described by (i) a particular combination $\mathcal{L}_i$ of enzymes that are available in that particular phase (along with their individual fixed concentrations), and (ii) its duration $\tau_i > 0$. Although replication is not strictly considered as a phase, it can similarly be represented by $(\mathcal{L}_r, \tau_r)$ with $\mathcal{L}_r = \emptyset$ and $\tau_r = 0$ because each replication is modeled as an instantaneous event where enzymes cannot perform any of their reactions.*

Following previous studies that used evolutionary optimization, I define mutation and recombination

operators for the individual instructions as follows:

- The concentration of a random enzyme can be decreased by up to two molecules. The concentration of a different random enzyme is correspondingly increased to keep the total number of enzymes constant.
- Up to two enzymes and their corresponding rewriting rules can be replaced by randomly picked different enzymes. The concentration variables and rate constants remain unchanged.
- In principle, the rate constants can be changed. However, I disabled this option in the simulations, as discussed below.
- The recombination (cross-over) operator builds a convex combination of two instructions $\mathcal{L}_1$ and $\mathcal{L}_2$ using the formula $\xi\mathcal{L}_1 + (1-\xi)\mathcal{L}_2$ with a randomly drawn weight $\xi \in [0,1]$.
- A second cross-over operator may similarly construct a combination of two already present enzymes and their associated rules, therefore adding a new enzyme in a non-random fashion. Due to the relative simplicity of the rules used in simulations, however, I did not use such operators here.
- If multiple instructions are used, their individual durations τ_i can be changed. For this, two randomly picked phases increase and decrease their individual durations by 10%, respectively (with respect to the time t_r between two replications). In addition, independently for each phase, enzyme availabilities (presence or absence, see Figure 4.4) of up to two random enzymes can be inverted.

The fitness of an instruction (or a schedule of instructions) is evaluated by comparing the patterns immediately before each of the cell divisions to the start pattern using the normalized Hamming distance. The initial patterns used here are compiled in Table 4.3. This yields $r+1$ distance values $d(i)$. For each start pattern and n_i independent starting point, I run the Gillespie simulation n_g times with different random number seeds. Then, I average the distance values over the Gillespie realizations, obtaining $1 - \langle d(i) \rangle$ as the autocorrelation function of the pattern. The fitness value is next computed as the sum of this autocorrelation function over the n_r cell divisions. It turns out that a simple hill-climbing approach is sufficient to obtain good solutions. Hence, a proposed mutation of the instruction (or schedule of instructions) is accepted if the estimated fitness increases. I stop the search if the best solution among all runs does not improve for n_i iterations.

Table 4.2: *Summary of the most relevant parameters for the EA. See text for details.*

Parameter	Value	Description
Specific to the EA		
n_i	1000	Stop criterion
n_s	10	No. of independent starting points / runs
n_p	1–4	No. of phases
Specific to the biology and the stochastic simulation		
n_r	50	No. of replications
n_n	30–150	Total no. of nucleosomes
c	Circular	Nucleosome organization
t_r	20	Time between two replications
n_g	20	No. of independent Gillespie realizations
n_a	5 or 10	Maximal no. of distinct active enzymes (for elementary and composite patterns, respectively; see Table 4.3)
n_b	1–5	No. of distinct chromatin domains per pattern (Table 4.3 and Figure 4.1)
k	30	Chromatin domain length (in nucleosomes)
n_m	10	No. of enzymes in the cell (per chromatin domain)
n_e	$n_b \times n_m$	Total no. of enzymes in the cell (all chromatin domains)
r_b	1	Binding reaction rate constant
r_d	5	Dissociation reaction rate constant

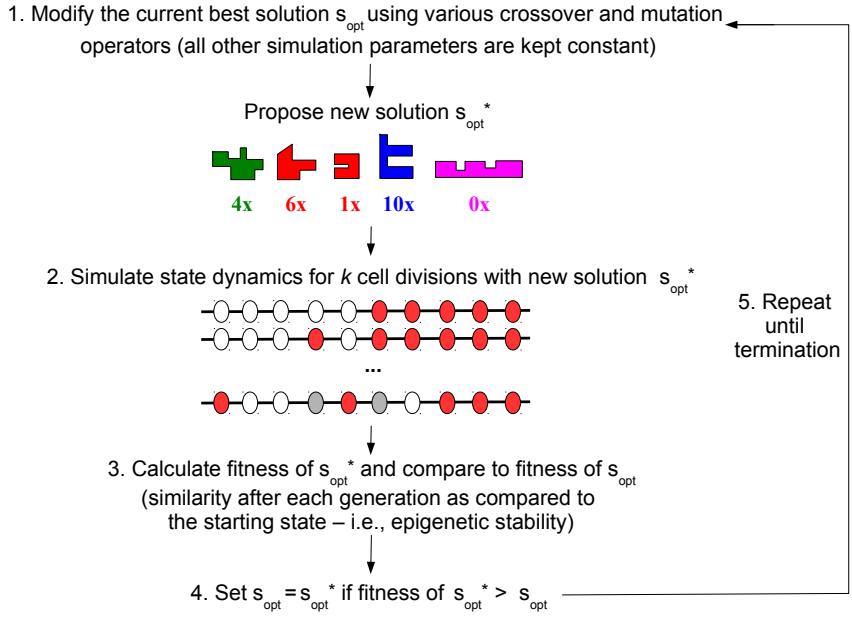


Figure 4.5: *Schematics of the main steps of the EA. See text for details.*

Table 4.3: *Summary of the start patterns used for the fitness evaluations. State 0 designates an unmodified nucleosome, whereas states 1 and 2 designate two distinct modified states of a particular histone PTM. The parameter k (here set to 30) is the length of each individual chromatin domain in each pattern, see Table 4.2 and Figure 4.1 for details.*

Pattern summary	Pattern length (n_n)	Pattern abbreviation
Elementary patterns		
1^k	30	1
0^k1^k	60	01
1^k2^k	60	12
Composite patterns		
$1^k2^k1^k$	90	121
$1^k2^k2^k$	90	122
$1^k2^k1^k2^k$	120	1212
$1^k2^k2^k2^k$	120	1222
$0^k1^k0^k$	90	010
$0^k1^k1^k$	90	011
$1^k0^k1^k$	90	101
$0^k1^k0^k1^k$	120	0101
$1^k0^k1^k1^k$	120	1011
$1^k0^k2^k$	90	102
$1^k2^k0^k$	90	120
$1^k2^k0^k0^k$	120	1200
$1^k2^k0^k2^k1^k$	150	12021

4.2.5 Simulations

For all patterns, I generated n_s initial instructions by randomly selecting n_l out of a total of 28 rewriting rules (i.e., the enzymes that implement them) as listed in Table 4.4. I also assigned the individual concentrations of these n_l enzymes randomly so that the total number of molecules equaled n_m. I kept all other parameters constant between the independent runs. Table 4.2 summarizes the parameters that I used for the EA and the individual stochastic simulations required for the fitness evaluation.

I simulated local chromatin state dynamics for a genomic region of 12 kb to 30 kb (depending on the pattern), a range that researchers also used in previous approaches [36] and that provides a good balance between computational speed and biological verisimilitude. I used a value of 50 for n_r because this reflects the Hayflick limit for how many times a cell can divide. Also, I used a circular nucleosome organization throughout the simulations to avoid artifacts for the boundary nucleosomes due to the rewriting rules and their neighbor-dependence. For a suitable number of enzymes available for the modeled chromatin region, I chose a value of 10 per chromatin domain (parameter n_m). Dependence on the number of chromatin domains is necessary to ensure comparability among

Table 4.4: *Summary of the 28 rewriting rules with patterns on $\mathcal{A} = \{0, 1, 2\}$ used for the simulations. For rewriting (RW) rules that are not intrinsically symmetric I also list their mirror image. The position in brackets is bound and modified, the two flanking positions remain invariant. The second column gives a rewriting rule abbreviation that will be used hereafter.*

RW rules realizing $0 \to 1$ reactions	α	**RW rules realizing $1 \to 0$ reactions**	γ
$[0] \to [1]$	α_1	$[1] \to [0]$	γ_1
$0[0]0 \to 0[1]0$	α_2	$0[1]0 \to 0[0]0$	γ_2
$1[0]1 \to 1[1]1$	α_3	$1[1]1 \to 1[0]1$	γ_3
$2[0]2 \to 2[1]2$	α_4	$2[1]2 \to 2[0]2$	γ_4
$0[0]1 \to 0[1]1$ or $1[0]0 \to 1[1]0$	α_5	$0[1]1 \to 0[0]1$ or $1[1]0 \to 1[0]0$	γ_5
$0[0]2 \to 0[1]2$ or $2[0]0 \to 2[1]0$	α_6	$0[1]2 \to 0[0]2$ or $2[1]0 \to 2[0]0$	γ_6
$1[0]2 \to 1[1]2$ or $2[0]1 \to 2[1]1$	α_7	$2[1]1 \to 2[0]1$ or $1[1]2 \to 1[0]2$	γ_7
RW rules realizing $0 \to 2$ reactions	β	**RW rules realizing $2 \to 0$ reactions**	δ
$[0] \to [2]$	β_1	$[2] \to [0]$	δ_1
$0[0]0 \to 0[2]0$	β_2	$0[2]0 \to 0[0]0$	δ_2
$1[0]1 \to 1[2]1$	β_3	$2[2]2 \to 2[0]2$	δ_3
$2[0]2 \to 2[2]2$	β_4	$1[2]1 \to 1[0]1$	δ_4
$0[0]1 \to 0[2]1$ or $1[0]0 \to 1[2]0$	β_5	$0[2]2 \to 0[0]2$ or $2[2]0 \to 2[0]0$	δ_5
$0[0]2 \to 0[2]2$ or $2[0]0 \to 2[2]0$	β_6	$0[2]1 \to 0[0]1$ or $1[2]0 \to 1[0]0$	δ_6
$2[0]1 \to 2[2]1$ or $1[0]2 \to 1[2]2$	β_7	$1[2]2 \to 1[0]2$ or $2[2]1 \to 2[0]1$	δ_7

patterns with varying length due to the concentration dependence of the enzyme reactions in the Gillespie algorithm. For the elementary patterns, up to five different enzymes may be selected, whereas I increased n_a to 10 for composite patterns. I found these values to be sufficient to obtain good solutions in the simulations. I chose the values for t_r, r_b, and r_d so that (i) enough reactions can take place between two subsequent replications, and (ii) the system tends to have free molecules available rather than a condition where all molecules are bound.

The unmodified state behaves differently from modifications since it is not distinguishable from erased information after replication. It may be helpful to allow more than one phase to obtain good solutions. Coordinated, phase-dependent enzyme availabilities may make it easier to systematically recompute the parental modification state. To identify the optimal number of phases n_p, I first ran the EA for the pattern `01` for solutions with one, two, three, and four phases and identified the solution with the highest score. I then ran all composite patterns that contain the unmodified state with only n_p phases (instead of variants with one, two, three, or four phases, respectively).

4.3 Results

Using the flexible software system that I developed (see end of results for details) to study the dynamics of histone PTM states, the difficulty of pattern reconstruction problem depends on the structure of the start pattern. Therefore, I initially summarize the observations for simple, elementary start patterns (Tables 4.5, 4.6 and Figure 4.6). The EA achieved stable solutions for constant patterns and patterns that consist of only modified nucleosomes with relative ease. It only rarely (or only in low concentrations) selected neighbor-independent rewriting rules because they easily introduce noise to the system. Similar to the results of Dodd et al. [38] and Hodges et al. [288], I found that chromatin domains can transiently multifurcate to form multiple smaller domains that remain stable for a particular amount of time, which was particularly pronounced for patterns that contain patches of unmodified nucleosomes (Figure 4.7). Noteworthy, I sometimes observed a gradual accumulation of errors during the lifetime of a cell (e.g., see the least stable solution for pattern 12 in Figure 4.6).

Constant patterns. As expected, it is trivial to find optimal solutions for the constant pattern 1 because the only rewriting rules required to recompute the parental pattern are either α_1 ($[0] \rightarrow [1]$) or α_5 ($0[0]1 \rightarrow 0[1]1$ or $1[0]0 \rightarrow 1[1]0$). These must be present in higher concentrations than rewriting rules that change 1 to 0 (class γ). Consequently, a large number of simulations achieved the optimal score. Also, contrary to other patterns, the inclusion of neighbor-independent rewriting rules, such as α_1, pose no disadvantage to the system.

Pattern 12. I also found that it is relatively easy to evolve a system that can stably maintain patterns when the parental nucleosome state consists of several chromatin domains of modified nucleosomes, as exemplified by patterns 12 and 121. I also found that four rewriting rules with approximately equal concentrations are sufficient for stable inheritance over 50 generations: α_3 ($1[0]1 \rightarrow 1[1]1$), α_5 ($0[0]1 \rightarrow 0[1]1$ or $1[0]0 \rightarrow 1[1]0$), β_4 ($2[0]2 \rightarrow 2[2]2$), and β_6 ($0[0]2 \rightarrow 0[2]2$ or $2[0]0 \rightarrow 2[2]0$). Notably, the boundary between differentially modified regions fluctuated stochastically because solely the available rewriting rules controlled it (Figure 4.6).

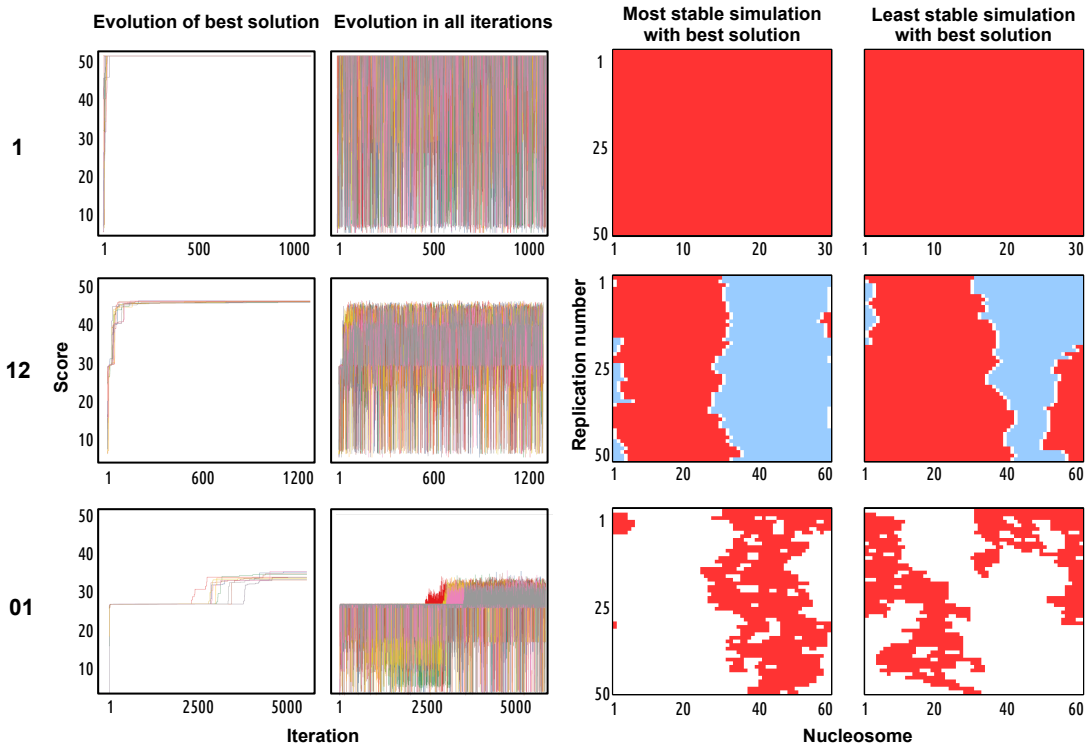


Figure 4.6: *Results from the EA for the three elementary patterns (top: 1, middle: 12, bottom: 01, 2 phases). The leftmost figure in each row shows the evolution of the best score (separately for each of the n_s independent starting points). The second leftmost figure shows the evolution of the score of the solution that the EA proposed in each particular iteration. The two rightmost figures in each row display the stochasticity of the best solution among all n_g independent stochastic simulations (left: most stable (highest score), right: least stable (lowest score)). For each visualization, the state of the system is shown directly before each replication (with the initial state at the top, and the state after the last replication at the bottom). The coloring of the different nucleosome states is analogous to the previous Figures (0: white, 1: red, 2: blue). For pattern 1, I detected no variation between the best and the worst performing run when I compared states only compared directly before each replication event.*

Pattern 01. It is substantially more difficult to find good solutions for patterns that contain a mixture of modified nucleosomes and unmodified nucleosomes such as 01, 101, and 102. The intuitive reason is that the 0-state in the target pattern is indistinguishable from the information lost during the replication event. Thus, the program of the CC lacks means to determine where states need to be regenerated and where the target state has already been reached. No good solutions seem to exist for this problem when only a single phase is allowed between replication events. To illustrate this, consider the pattern 0^k1^k. Rewriting rules are needed to re-establish the right part of the pattern that contains the nucleosomes in state 1 after each cell division. These can principally be constructed in two ways: (i) The rewriting rule α_3 ($1[0]1 \rightarrow 1[1]1$) ensures that it can only be applied in the right part. However, having only this rewriting rule is too strict because particular patterns, such as 01001, cannot be repaired otherwise. This would be possible with (ii)

α_5 ($0[0]1 \rightarrow 0[1]1$ or $1[0]0 \rightarrow 1[1]0$); however, this is also applicable at the boundary of the 0 and 1 region and can therefore slowly spread into the 0-region, leading to gradual loss of the ancestral signal. Rewriting rules that do not incorporate the state of neighboring nucleosomes (e.g., α_1 ($[0] \rightarrow [1]$)) are also not helpful in this regard, since they introduce additional noise to the system.

Admitting more phases, i.e., variation of the gene expression program through the cell cycle, can mitigate this difficulty. Different phases can then serve a particular purpose and, collectively, they aim at restoring the parental pattern. In practice, I found that solutions with more than one phase become only marginally better with two or more phases, possibly due to the greatly expanded parameter space of the solution (Table 4.5). It may therefore be necessary to run the EA for a substantially longer time to obtain good solutions that consist of more than one phase. Nevertheless, the stability of the best solutions I found (independent of the number of phases) is not comparable to solutions for patterns 1 and 12. This is indicated by both the attainable scores and by the time course visualization of the patterns in Figure 4.6.

Composite patterns. For the composite patterns, the results were as follows: Patterns that are combinations of all three elementary patterns (i.e., with 0-1, 0-2, and 1-2 chromatin domain transitions) produced the lowest scores, despite allowing the EA to increase the maximal number of active enzymes (see below). Variations of the pattern 12 produced the best scores. For the pattern 12021 and 1011, the simulations consistently lost the 0 patch in the middle of the patterns, even in the best solutions.

I also found that solutions for the three elementary patterns can be used with comparable quality for repetitions of elementary patterns. A good example is the pattern 12, where all of the variations tested produced scores that differed by less than 10% from the original score. This occurred even though the patterns were either more complex (121, 1212) or contained chromatin domains of unequal length (122, 1222). For the pattern 01, I observed a similar result with the patterns 010, 011, 101, 0101, and 1011 although the scores were up to 20% smaller (data not shown). In all cases, the score may be improved by adjusting enzyme concentrations and phase durations, particularly for patterns with a different length than the original pattern.

However, I found that individual solutions from elementary patterns cannot merely be combined for a more complex pattern, particularly with multiple phases. For example, the combination of the individual solutions for the patterns 01, 02, and 12 does not produce a good result for patterns that contain all these three types of transitions such as 102. The combination of different solutions — i.e., the simultaneous presence of more enzymes — apparently interferes with the "strategy" of the partial solutions: The additional enzymes act at the newly produced modifications and obliterate the nascent pattern.

I also investigated whether the performance could be improved by optimizing the rate constants for binding and/or dissociation rates, along with the rewriting rules themselves. Although the scores

Table 4.5: *Summary of the results from the EA. For each pattern, I give the number of iterations $n_{\mathbf{it}}$ after which the EA finished, as well as the score of the best solution, the number of phases n_p that solution consisted of, and the number of distinct enzymes that the EA selected (n_a). For more details on the composition of the solutions, see Table 4.6.*

Pattern	n_p	$n_{\mathbf{it}}$	n_a	Best score
Elementary patterns				
1	1	1001	4	51
12	1	1180	5	45.6
01	1	4083	5	35.1
01	2	5360	3	36.6
01	3	2259	5	35.5
01	4	4823	5	35.6
Composite patterns				
121	1	2717	4	47.8
122	1	1346	6	44.7
1212	1	1367	4	45.0
1222	1	3595	4	48.4
010	2	5229	7	43.2
011	2	6574	6	36.6
101	2	5834	5	35.0
0101	2	5106	6	35.9
1011	2	1002	5	38.4
102	2	5116	5	33.4
120	2	3840	6	33.2
1200	2	5231	6	34.4
12021	2	3429	7	39.1

were comparable, I obtained solutions that are much more tailored towards the reconstruction of a particular initial pattern length (data not shown). This is because the reaction rate constants also control how many reactions may take place during a particular time in the stochastic simulation. By allowing them to vary, the enzymes and the corresponding reaction rates become tailored to the specific pattern length. The EA then also more frequently selected neighbor-independent enzymes because their frequency of selection can be controlled by the reaction rates.

Finally, I examined to what extent the parameter n_a, which limits the maximal number of active enzymes, has an influence on the quality of the solutions. I found that selection sometimes tends to increase the number of enzymes by including rewriting rules that are rarely applicable. This was particularly true for the pattern 01, where the score for the best solution with one phase increased to that of solutions with multiple phases using the original value of n_a (data not shown). This appears to be a means of adjusting reaction rates to decrease the number of reactions that take place between replications.

To verify that the best solutions for the different patterns are not specific to the parametrization of

Table 4.6: *Summary of the best simulation for each of the elementary patterns. The second column presents the number of phases n_p that compose the solution. The third column summarizes the best solution and lists the enzymes active in a particular phase, together with their abundance (in brackets). For the patterns 1 and 01 with a single phase, I found many distinct optimal solutions (see text), and only one representative is included here. For the pattern 01, I performed multiple independent evolutionary optimizations with a different number of phases (see text). If I set multiple phases, the individual phase durations are also presented (in percent).*

Pattern	n_p	Best solution $\tau_i, \mathcal{L}_i$
1	1	τ_1(100%), $\mathcal{L}_1$:$\alpha_3(3)$, $\alpha_5(4)$, $\alpha_7(1)$,$\delta_2(2)$
12	1	τ_1(100%), $\mathcal{L}_1$:$\alpha_3(3)$, $\alpha_5(5)$, $\beta_4(3)$, $\beta_6(5)$, $\gamma_7(4)$
01	1	τ_1(100%), $\mathcal{L}_1$:$\alpha_3(2)$, $\alpha_5(8)$, $\alpha_7(2)$,$\delta_4(4)$
	2	τ_1(17.8%), $\mathcal{L}_1$:$\alpha_3(6)$,$\alpha_5(1)$,$\delta_5(13)$ τ_2 (82.2%), $\mathcal{L}_2$:$\alpha_3(6)$
	3	τ_1(45.5%), $\mathcal{L}_1$: see phase 3, + $\alpha_5(7)$ τ_2(31.8%), $\mathcal{L}_2$: see phase 1 τ_3(22.8%), $\mathcal{L}_3$:$\alpha_3(4)$,$\alpha_6(4)$,$\alpha_7(3)$,$\delta_4(2)$
	4	τ_1(35%), $\mathcal{L}_1$:$\alpha_5(7)$,$\alpha_7(4)$,$\beta_5(4)$ τ_2(35%), $\mathcal{L}_2$:$\alpha_5(7)$,$\alpha_7(4)$,$\delta_4(3)$ τ_3(15%), $\mathcal{L}_3$:$\alpha_5(7)$,$\alpha_7(4)$,$\beta_5(4)$,$\delta_4(3)$ τ_4(15%), $\mathcal{L}_4$:$\alpha_3(2)$,$\alpha_7(4)$,$\delta_4(3)$

the model and the specific pattern length, I also tested the sensitivity to parameter variations (Table 4.7). Specifically, I tested the effect of a linear nucleosome string rather than a circular one, the time between two replications, the number of nucleosomes and replications, and the dissociation rate of all enzymes. In summary, I found that the solutions produce very similar scores in most of the parameter space for the patterns 1 and 12, whereas for the pattern 01 and its compositions the best solutions strongly depended on the kinetic parameters. I also found that the number of nucleosomes must not be too low (a value around 40 sufficed for robustness). Otherwise, stochastic effects may irreversibly destroy the parental signal. Similarly, the time t_r between two replications must be long enough to allow for recomputation of the parental pattern. Some noteworthy effects that I observed while varying the parameters are summarized in Figure 4.7.

The source code of a C implementation of the software system can be obtained under the GNU Public License from `http://www.bioinf.uni-leipzig.de/Software/StoChDyn` and consists of two separate programs: the stochastic simulation of the dynamics of histone PTMs using Gillespie's approach (*StoChDyn*) and the EA (*Evo-ES*) that uses *StoChDyn* to evaluate its solutions.

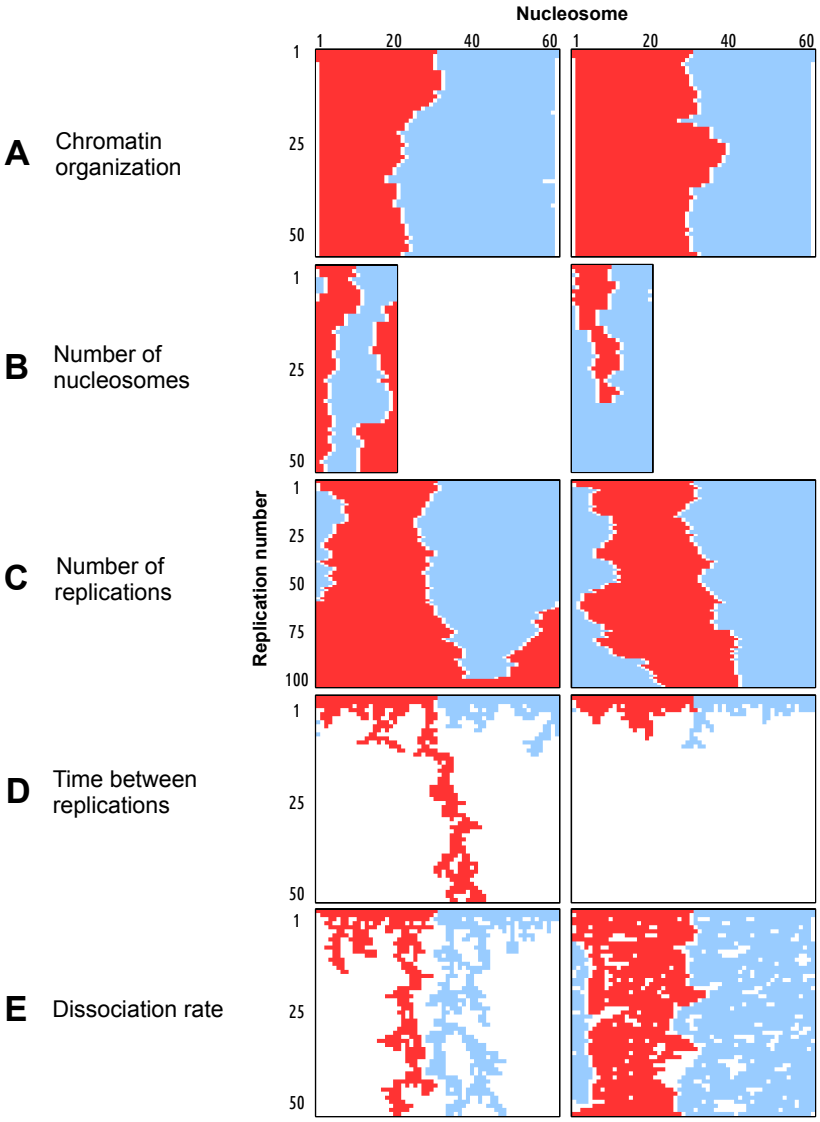


Figure 4.7: *Visualizations from selected simulations for pattern 12 of the robustness analysis. One distinct parameter is fluctuated in each row (A–E, as indicated on the left) as compared to the best solution for this pattern, and two exemplary runs of the same simulation (except for E) are shown that highlight some noteworthy effects that we observed. The coloring is analogous to Figure 4.6.*
(A) Variations in the chromatin organization at the boundaries (here: linear instead of circular).
(B) Variations in the number of nucleosomes (here: 20 instead of 60), illustrating that reducing the number of nucleosomes increases the likelihood that the signal is lost due to stochasticity.
(C) Variations in the number of replications (here: 100 instead of 50), showing that chromatin domains can gradually disappear (left) or change their exact location due to stochasticity (right).
(D) Variations in the time between replications (here: 0.5 instead of 20), highlighting that the original signal may gradually get lost if the cell has not enough time to recompute the parental pattern.
(E) Variations in the dissociation rate of the enzymes (here: 0.05 (left) and 0.15 (right), respectively, instead of 1), illustrating that the enzyme kinetics may also have a large effect on pattern stability. I did not modify the binding rate constants; however, I reduced the dissociation rate constants, which means that enzymes need more time to perform their designated reaction.

Table 4.7: *Summary of the robustness analyses for the best solution for each elementary pattern. The table lists the parameter value intervals when the score of the modified solution (with one changed parameter) achieved more than 90% of the score of the original best solution. The variable s denotes the step size. When I varied the number of replication n_r, I calculated the ratio of the score and the corresponding maximal possible score and compared the two ratios using the 90% threshold for better comparability because n_r has a direct influence on the maximal possible score.*

Parameter	Parameter values			Pattern		
	Original	New, varied		1	12	01
Nucleosome organization (c)	Circular	Linear		All	All	All
Number of nucleosomes (n_n)	60	10–200	($s = 10$)	All	> 40	60–90,140
Number of replications (n_r)	50	5–100	($s = 5$)	All	All	< 55
Time between replications (t_r)	20	0.5–40	(s variable)	> 5	> 1	19–20
Dissociation rate (r_d)	5	0.05–100	(s variable)	> 0.5	> 0.15	4

4.4 Discussion

I have queried whether the propagation of patterns of histone PTMs across cell divisions can be seen as a computational problem and if so, whether chromatin is organized in a way that is amenable to the solution of this problem. The answer is twice affirmative. I demonstrated that the faithful propagation of patterns of histone PTMs can be interpreted as a computational problem that is achievable through a small collection of rewriting rules. These rewriting rules are abstractions of a well-described class of enzymes and enzyme complexes combining reader, writer, and eraser domains for specific histone PTMs. For the best solutions, the EA selected almost exclusively enzymes that are dependent on the states of neighboring nucleosomes (except for the trivial pattern 01). This highlights that context-dependency is crucial for such inheritance systems because context-independent enzymes easily introduce too much noise to the system that further complicate the recomputation of parental state patterns after cell division. Indeed, for many histone-modifying enzymes, it is well-known that their binding affinities are highly influenced by the presence or absence of particular histone PTMs or other signals nearby.

This modification process is intrinsically stochastic and crucially depends on concentrations of the available enzymes and histone PTMs. As a practical implementation, I employ a detailed stochastic simulation of chromatin state dynamics to approximate the physico-chemical constraints of my approach. Here, the probability for applying rewriting rules is modeled explicitly in reaction rates for binding and dissociation following the laws of mass action.

I showed that stable propagation of complex patterns without the need for explicit boundary elements is possible in the model although not for all types of patterns. This is consistent with the findings of Hodges et al. [288] who recently proposed that explicit boundary elements may not be required for H3K9me3 domains because their size and propagation may be naturally limited by chromatin remodeling processes. Nevertheless, the maintenance of (approximate) boundaries

between differentially modified chromatin domains is a main challenge of any epigenetic inheritance mechanism. In practice, additional security measures that prevent spreading may furthermore be required for faithful long-term inheritance. Indeed, several mechanisms that prevent spreading may trivially solve the problem of restoring the parental modification pattern when exact boundaries are important. Examples include nucleosome-depleted regions (for example, see [361–363]), bound proteins (e.g., CTCF) or histone variants (e.g., H2A.Z) in the vicinity of positions that border the different chromatin domains, ncRNAs [208], or the marking with a particular histone PTM directly after replication due to other signaling factors.

In the model presented here, long-range interactions are not necessarily required for stability, which contrasts to what is argued by Dodd et al. [36]. However, the presence or absence of long-range interactions has no impact on the computational power of the CC or the conclusions with respect to epigenetic inheritance. Generally, epigenetic inheritance should be seen as an ensemble of different strategies that collectively aim to transmit a particular chromatin state throughout cell division. Patterns of histone PTMs fit this bill. The propagation of one out of several alternative modified states provides the required memory across divisions. Nevertheless, only a few primary histone PTMs that typically form large and homogeneous chromatin domains (e.g., H3K9me3) may be copied in a self-propagating manner as described here. The transmission or recomputation of other more secondary histone PTMs, however, likely depends on other factors [178]. On the other hand, the inheritance of promoter-specific modifications (which typically cover only a few nucleosomes) is likely implemented in a different way than the propagation of large homogeneous chromatin domains because short domains are much more difficult to inherit due to pure stochasticity (e.g., see Figure 4.7 B).

Pattern stability is influenced by a number of factors such as dynamic chromatin remodeling events, the up- and downregulation of genes that code for or regulate the corresponding histone-modifying enzymes, pattern complexity [39], and the length of the state to be maintained. In summary, patterns of histone PTMs are often an ongoing enzymatic competition between their placement and removal. Altering this steady-state balance pushes either towards the accumulation of the mark or its erasure [288]. This is apparent most clearly in embryonic reprogramming [338, 364].

The finding that patterns containing patches of unmodified nucleosomes are more difficult to inherit than modified ones (irrespective of the number of phases) due to the ambiguity of the unmodified state raises the question of biologically relevance. Due to the sheer complexity of histone modifications, the vast majority of nucleosomes may carry at least one modification, which could facilitate recomputation of the parental patterns and resolve the difficulty of stably inheriting such domains. Additionally, other chemical signals within the vicinity of a nucleosome, such as DNA methylation, the presence of histone variants or spatial contacts with genomic loci, and protein complexes that are themselves retained through cell division, may be specifically used to backup the information of nucleosome left unmodified intentionally.

I emphasize that the focus in this thesis is on the computational task of re-constructing complex patterns of histone PTMs that is typical for somatic cells. I do not claim that epigenetic inheritance across the germ line follows the same paradigm. Information inherited through the germline for an effectively infinite number of generations is subject to Eigen's error threshold [365], which links the amount of stably inheritable information to the accuracy of information propagation. Whereas effective proofreading mechanisms limit replication errors to a single mutation per round of replication for genomic DNA, no mechanism is conceivable that would achieve a similar accuracy for histone PTMs. As a consequence, the amount of stably inheritable epigenetic information is severely limited. Consistent with this theory, most, if not all, of the extraneous epigenetic information is erased during spermatogenesis and oogenesis. The resulting totipotent state is characterized by global erasure of DNA methylation, chromatin reorganization, differential regulation of histone-modifying enzymes (e.g., the tendency for the upregulation of histone de-modifying and downregulation of histone-modifying enzymes) [366]. The initial stages of embryogenesis are governed by a gene regulatory network dominated by transcription factors (e.g., reviewed in [367]), partial ejection of nucleosomes [368] and therefore a reduction in the availability of a major epigenetic information carrier. Indeed, it seems that only few epigenetic modifications are part of the epigenomic basal state (e.g., strong heterochromatin formation of genes linked to differentiation [368] or imprinting and poised promoters). In contrast, the error threshold does not preclude inheritance of complex patterns of histone marks in somatic cell lines because the number of generations is limited, and usually small. Here, the degradation of the epigenetic information is acceptable for a while but inevitably leads to daughter cells whose epigenetic patterns are damaged beyond repair. This effect may thus constitute an epigenetic version of aging. Intriguingly, histone PTMs may also function as molecular timers, as evidenced by H3K79me that couples cell cycle progression to changes in the epigenome [170].

The model contains a number of biological parameters. However, experimentally validated and reliable values are only available for a few of them (such as the number of DNA replications and typical chromatin domain lengths). For other parameters, such as the various rate constants for binding and dissociation reactions and the number of active enzymes, reported values span several orders of magnitude (see [289] for an excellent overview on the subject). For example, rate constants for chromatin-associated proteins have strikingly different kinetics, with average residence times ranging from a few milliseconds for some TFs and remodelers up to several hours for nucleosome components [289]. However, until recently, different methods to determine these average residence times of TFs on chromatin such as FRAP (fluorescence recovery after photobleaching) yielded highly unequal results, with differences of up to three orders of magnitude (reviewed in [369]). Newer methods, such as SMT (single molecule tracking) and improvements of the FRAP method corrections, now finally seem to reach a consensus [369]. Likewise, the concentrations of various chromatin-associated proteins and binding affinities also vary greatly (reviewed in [289]).
In terms of total numbers of enzymes present and active for a particular genomic region, Steensel

[80] provided some estimates and calculated that each nucleosome may be in contact with up to 30 proteins (see [80] for more details), which is much higher than the values I used in the simulations (Table 4.2). However, their individual residence times as well as the binding influence with respect to other proteins remains unclear. Lastly, as mentioned in Section 2.3.3, Hathaway et al. [277] estimated that H3K9me3 domains spread with ≈0.18 nucleosomes per hour up to a length of ≈10 kb. The authors also found that the observed domain lengths are only compatible with their model if the relative propagation rate does not exceed a particular threshold.
All of these estimates theoretically allow to adjust the simulation parameters accordingly. Thus, in summary, for a specific set of enzymes and histone PTMs, it may be possible to run the chromatin model primarily with biologically validated and plausible parameters in the near future.

The chromatin model and implementation can furthermore be extended in various directions. First, to reduce computing time and to allow the modeling of larger genomic regions, approximative stochastic simulation algorithm variants as well as the faster but still exact next reaction method as proposed by Gibson et al. [352] may be implemented. Additionally, the efficiency of the EA may be improved by including ant colony optimization algorithms (reviewed in [370, 371]), for example. Although it is not entirely clear if such algorithms indeed find better solutions, they seem to be particularly worthwhile for solutions with multiple phases due to the greatly increased search space. Second, additional components may be added to the system to account for biological complexity and realism such as different DNA replication and histone segregation models, the explicit modeling of nucleosomes as an octamer (i.e., modeling that each histone is present twice), different models for the retainment of parental histones after DNA replication, and the inclusion of histone turnover. The latter is particularly relevant because it seems to be a necessary component in real systems, as exemplified by H3K9me3 domains [277, 288]. However, the modeling and applicability of some additional phenomena is somewhat limited by the insufficient current biological knowledge and therefore awaits further research. An extension of the phases concept that allows phase-dependent concentration of enzymes rather than their mere presence or absence would also be worthwhile. Alternatively, this could be achieved by the inclusion of "events" (i.e., specific time points when, for example, the concentration of a particular enzyme changes). These extensions would allow to investigate what effect abrupt or gradual changes in the concentration of particular enzymes have for epigenetic stability, and how vulnerable the system is to small perturbations of the enzyme concentrations (although I performed the latter already to some extent, data not shown), for example. A third extension is to model the dynamics of more than one histone. Although this is already possible in the implementation, I never explicitly modeled this. As shown by Sneppen et al. [39], the combination of two distinct histone PTMs creates large numbers of different circuits capable of heritable bistability, and it would be interesting to study the results of the EA.

Part II:

CUSTOM EXPRESSION MICROARRAY DESIGN AND THE SIGNIFICANCE OF THE CHROMATIN COMPUTER IN ALZHEIMER'S DISEASE

Chapter 5

Designing Custom Expression Microarrays in the Post-ENCODE Era

5.1 Motivation and Background

Microarrays are a powerful technology for genome-wide transcriptome profiling that have been used ubiquitously in biomedical research for almost two decades now. They allow quantifying the expression of thousands of nucleic acid samples (RNA transcripts or target sequences) by hybridizing them to known sequences (probes) in a massively parallel and often genome-wide manner in a single experiment. Probes are typically 25–60 bp oligonucleotides long, and in essence, they are immobilized on a solid surface and bind to complementary targets (hybridization) [372]. Microarrays can be applied in many different areas such as comparative genomic hybridization/copy-number analysis (CGH microarrays), single nucleotide polymorphisms detection (SNP microarrays), chromatin immunoprecipitation (ChIP-on-chip microarrays), gene expression and microRNA profiling (gene expression microarrays), and DNA methylation (methylation microarrays).

Most microarray providers offer custom expression microarrays (CEMs) for which the represented target sequences can be precisely defined. CEMs are increasingly popular because they are more cost-effective than tiling arrays and offer more flexibility. They have been used frequently to address a variety of questions such as the development of genetic markers [373] and the generation of gene models for particular treatments [374] to improve understanding of the molecular mechanisms [375] or the transcriptome response to different external signals [376].

The workflow of custom expression microarray design can broadly be divided into three parts: target sequences selection, probe design, and probe selection (Figure 5.1). Probe design is important and a high-quality oligonucleotide probe must be (i) sensitive, (ii) specific, and (iii) isothermal with the other probes [377]. Although numerous tools exist for designing probes for a set of RNA targets, they only address the first and third criteria accordingly. The measures taken to ensure probe specificity are often too simplistic. For example, available probe designers only find probes that are unique with respect to currently annotated transcripts. However, such a strategy is insufficient because RNA samples may contain novel transcripts not yet contained in any public database. Cross-hybridization

must therefore, in theory, be tested against the complete (unknown) transcriptome in order to not overestimate probe specificity. This is particularly important in light of the growing importance of ncRNAs and chromatin-associated RNAs in particular (see Section 2.1.4.3).
Vertebrate genome complexity is indeed often neglected and underestimated by existing tools although the large majority of the genome is capable of transcription in a highly time-, tissue-, and developmental-specific manner (see Section 2.1.3). Genomic loci often encode a variety of transcripts comprising several isoforms of protein-coding and non-coding RNAs with a large number of (partly) overlapping transcripts. A careful selection of specific regions in target sequences in which probes should be placed is therefore necessary. If probes are designed independently for each of these overlapping transcripts, a probe may be located in exons that are shared between isoforms (transcripts) of the same (different) gene(s). Specificity of those probes is then low due to possible cross-hybridization with different RNAs transcribed from the same genomic locus. Probes should therefore be placed in exons or splice-sites that are unique to a target sequence.

In addition, if multiple datasets are integrated, the selection, unification, and generally preprocessing of heterogeneous target sequences is time-consuming, error-prone, and non-trivial. Surprisingly, however, this has yet to be addressed in a systematic and thorough fashion [378]. For example, none of the existing tools provide a framework to address the following "design strategies" before and after probe design:

- How can overlapping target sequences be handled and what individual advantages and disadvantages do different approaches have? How can cross-hybridization be minimized?
- Which strategies exist to efficiently use the limited available space on the CEM? Which criteria may be used to select among a set of target sequences?
- How can target sequences be processed so that flexible strategies with respect to the number of probes per target sequence can be realized?
- Which measures may be useful to evaluate the success of target selection and subsequent probe design? In particular, what is the coverage of target sequences with probes?

All of these issues may substantially influence the reliability, accuracy, and interpretability of the resulting expression measurements [379]. Indeed, high-quality microarrays are fundamental for any meaningful interpretation of the data. As we will see in Chapter 6, the development of a pipeline that aims in the production of high-quality CEMs is therefore an important prerequisite for addressing biologically relevant questions such as the significance of chromatin and the chromatin computer in Alzheimer's disease (AD).

5.2 Methods and Results

5.2.1 The Custom Array Design Pipeline

I developed a bioinformatics pipeline (CAD pipeline) that addresses the first and last step in a typical custom expression microarray design workflow (i.e., suitable selection of target sequences and probes) more profoundly than available tools (Figure 5.1). Specifically, it may be used as an automated tool assisting in the error-prone and time-consuming steps of generating a set of unified target sequences suitable for probe design out of a set of (heterogeneous) input datasets (target sequences selection), given specific design strategies with respect to the representation of complex transcriptional loci (see Section 2.1.3) and user-defined, dataset-specific parameters. After subsequent probe design using external programs, the CAD pipeline may also be used for probe selection by rigorously evaluating their specificity, discarding non-specific probes, and calculating various statistics that evaluate to what extent the target sequences are represented by high-quality probes.

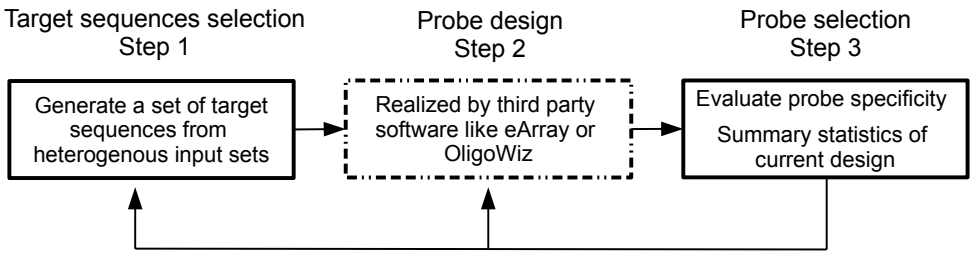


Figure 5.1: *Schematics of the concept and positioning of the CAD pipeline in the workflow of CEM design. The CAD pipeline specifically addresses the target and probe selection but leaves the probe design to external tools. See text for details.*

The CAD pipeline integrates particularly well with the eArray platform from *Agilent* but also supports other probe design software due to the use of standard file formats. A detailed schematic overview of all its steps (and the corresponding web server, see Section 5.2.2) is presented in Appendix B. In the following, the main features of the pipeline are described.

5.2.1.1 Dataset Unification

Dataset unification specifically addresses the conversion of all datasets to one format with identical genome assembly versions. The CAD pipeline accounts for the variety of known classes of RNA in complex genomes [380] by allowing the incorporation of an arbitrary number of input files (therefore collectively representing, for example, small ncRNAs, protein-coding genes, and long ncRNAs).

This is particularly useful if heterogeneous datasets from different public databases are integrated because dataset unification may be time-consuming and error-prone. In the CAD pipeline, both sequence-based as well as position-based formats (FASTA and BED, respectively) are supported. Whereas position-based formats may require a conversion of the genomic coordinates to a particular target genome assembly version, sequence-based formats necessarily require a mapping step. For mapping genomic coordinates from one assembly to another one (e.g., from hg18 to hg19), I use the program liftOver[1]. The mapping of sequences is performed using BLAT[2] [381], and its parameters may be arbitrarily customized. Non-uniquely mapping target sequences are discarded due to their inherent ambiguity. If sequences map to multiple spatially separated blocks or regions (e.g., two exons interspersed by an intron), each block is subsequently treated as a separate target sequence.

For some target sequences (reliable) strand information may not be available, and it may therefore be useful to additionally include the sequences from the complementary strand. In the CAD pipeline, this option can be enabled file-specifically and therefore provides maximum flexibility. Examples are datasets containing genomic regions that are known to harbor ncRNAs with conserved secondary structures based on structure prediction tools such as Evofold [382] and RNAz [383, 384] or datasets derived from chromatin immunoprecipitation (ChIP) experiments or any of its variants (e.g., see [188]). Additionally, antisense sequences may be deliberately included to investigate if sense transcripts are regulated by their corresponding antisense transcripts.

In the CAD pipeline, the automatization is realized through a central configuration file that summarizes all datasets and their specific parameters such as file format (e.g., FASTA or BED), genome assembly version (e.g., hg18), the various filters that should be applied, specific mapping options (if applicable), and if sequences on the complementary strand should also be included. Furthermore, global parameters are defined such as the specific strategy to handle overlapping target sequences, the target genome assembly version (e.g., hg19), and various filter-related parameters (see the following pages for details).

5.2.1.2 Target Sequence Filters

The CAD pipeline provides a set of sequence filters to discard (parts of) target sequences that are redundant or negligible for the purpose of the study. Additionally, they may also be used to reduce the number of target sequences if the array capacity is reached for speeding up the probe design process or to improve interpretability of the expression results (see below for an example). For maximum flexibility the following filters may be applied individually for each input dataset:

- Length filter: A length filter may discard target sequences that are too short or, optionally,

[1] available at http://genome.ucsc.edu, last accessed in August 2013
[2] available at http://genome.ucsc.edu, last accessed in August 2013

too long. This is useful for target sequences shorter than the actual probe length, for example.

- Redundancy filter: Target sequences from different input datasets may be redundant (i.e., identical start and end positions), particularly if they originate from publicly available databases. Thus, only one representative target sequence is retained that combines the individual annotations of all discarded ones.

- Negative set filter: The negative set filter specifies genomic loci that should not be included for probe design (e.g., repeat regions, UTRs, coding exons). Application of this filter is useful, for example, in the following scenarios:

  – The number of target sequences must be reduced.

  – The study focuses primarily on a particular set of transcripts such as ncRNAs. For ensuring that non-coding and coding transcripts can be distinguished, all coding parts may be eliminated to design probes for the non-coding part of the sequences only (Figure 5.2). Similarly, this strategy is useful for alternative splicing forms of coding and non-coding sequences because probes in overlapping parts can also not distinguish between coding and non-coding variants.

  However, the application of this filter may (i) split a particular target sequence in multiple target sequences if one or more parts in the middle of the target sequence overlaps with regions that should be excluded, (ii) remove a target sequence altogether if it overlaps completely with an excluded region, and (iii) retains only such a small fraction of the target sequence that the length filter takes effect.

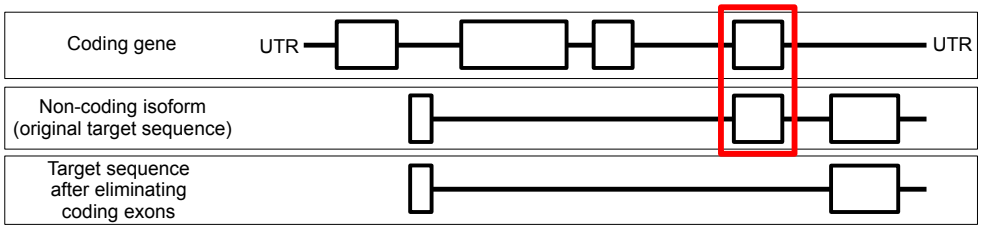


Figure 5.2: *Illustration of the negative set filter. An example for the application of the negative filter is shown for a non-coding isoform (target sequence) of a particular gene that is subject to splicing. Only the elimination of the coding sequences (exons) in between ensures that a distinction between non-coding and coding transcripts is possible on the probe level. For more details, see text.*

5.2.1.3 Handling of Overlapping Sequences

If target sequences from several sources are integrated, an important design decision is how to handle overlapping target sequences. This is particularly relevant because one the one hand,

genomic loci often encode a variety of transcripts comprising several isoforms of protein- and non-coding transcripts (see Section 2.1.3 and [385]), and on the other hand, sequences from publicly available databases also show a large amount of overlap. The mapping of a probe to a particular transcript/target sequence should be as unique as possible, and sequence redundancy should be minimized.

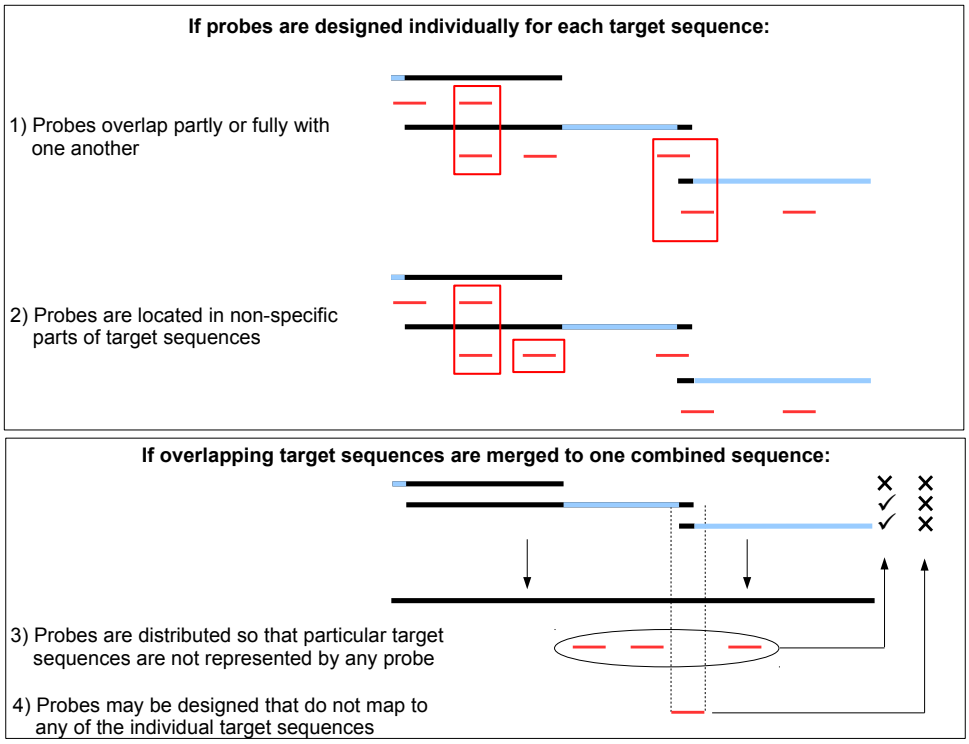


Figure 5.3: *Issues introduced by overlapping target sequences with regard to subsequent probe design and probe distribution. A schematic depiction of four issues introduced by overlapping target sequences with respect to subsequent probe design and probe distribution are shown. Regions marked in blue are not shared by any of the other target sequences and are therefore "specific" or unique for the particular target sequence. Designed probes are colored in red. For more details, see text.*

Overlap of target sequences introduces at least four issues with regard to the distribution and location of probes (see also Figure 5.3):

- Probes from overlapping target sequences may also overlap, thus causing potential cross-hybridization problems (non-specificity).
- Probes may be located in parts of target sequences that are shared by at least one other target sequence. Therefore, because the part is not unique to the sequence (Figure 5.4) it is inherently impossible to map this probe to a unique transcript, causing ambiguity with respect to from which transcript the expression signal originates.

- If overlapping target sequences are merged to one combined sequence (second strategy, see below), the following two additional peculiarities may arise:
  - If probes have been designed for a set of partly overlapping alternative transcripts for a particular gene, for example, they may be distributed so that individual transcripts are not represented by any probe, which renders their expression measurements impossible.
  - Probes may be designed that do not map to any of the individual target sequences because they are located at the boundaries of two adjacent or overlapping target sequences.

Therefore, sequence overlap may have major consequences for the interpretability of expression results and has, to the best of my knowledge, not yet been addressed explicitly and systematically. A specific strategy should thus be well-considered. In the CAD pipeline, the following three distinct strategies to handle overlapping target sequences can be selected, each of which has its own advantages and disadvantages (Table 5.1 and Figure 5.4):

1. **Do not merge overlapping sequences and ignore overlap among target sequences**
   Each target sequence is treated individually irrespective of any sequence overlaps. The main advantage of this approach is that probes can subsequently be specifically designed for each sequence. However, they may substantially overlap, they may be rarely located in parts that are not shared by any other target sequence, and the maximum number of designed probes may be (too) large, thus potentially wasting array capacity due to probe redundancy.
2. **If target sequences overlap, merge them to one combined sequence**
   All overlapping target sequences are merged to one combined sequence, therefore eliminating any overlap among target sequences. Thus, probes will also not overlap (unless specified explicitly) and the number of designed probes is, in general, smaller compared to the first strategy (Table 5.1). However, probe specificity and individual target sequence coverage may be low. This strategy is favorable if genes should be represented on the microarray but a distinction into different isoforms is irrelevant.
3. **If target sequences overlap, use only the non-overlapping specific regions of each target sequence**
   Only the part(s) specific for each target sequence (i.e., the part(s) not overlapping with any other target sequence) are retained, and all unspecific parts are eliminated. Although this strategy results in non-overlapping probes with high specificity, sequence coverage may be very low, particularly if only a small percentage of target sequences have any specific sequence parts of sufficient length so that probes can in fact be designed. This strategy allows a distinct quantification of the expression of various isoforms for a particular loci.

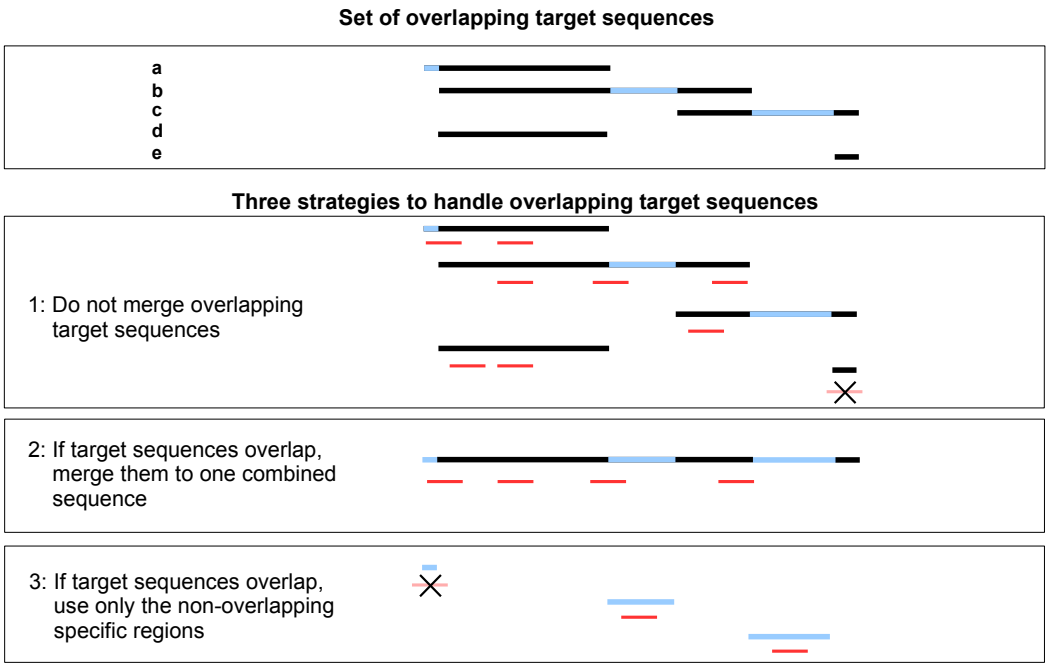


Figure 5.4: *Schematic overview of three different strategies to handle overlapping target sequences. The three different strategies are applied to a set of five overlapping target sequences for which probes should be designed (labeled a–e). Note that for some target sequences or parts thereof, probes cannot be designed because the sequence length is shorter than the length of the actual probes (indicated by x). The color scheme is analogous to Figure 5.3. For more details, see text.*

5.2.1.4 Number of Probes per Target Sequence

Expression measurements from sequences with only a single probe may yield substantial signal differences [386, 387]. For example, Chou et al. [387] examined the effect of this so-called measurement bias and found a strong dependence with the number of probes per target sequence (hereafter called n_p). If n_p is very low, measurement bias may be very high, whereas the bias decreases non-linearly if n_p is increased. To tackle this problem, one can use large n_p values to obtain more accurate expression measurements. However, due to the limited space on the microarray, n_p must not be too large. Furthermore, using a fixed value for n_p cannot adequately represent all target sequences due to their length differences that are frequently large.

While the number of probes per target sequence is only relevant for probe design itself, which is not integrated into the CAD pipeline, I implemented a very flexible way of preprocessing target sequences that easily allows to employ different strategies in the probe design with respect to the number of probes per target sequence. The approach presented here is to partition target sequences into different distinct sets according to their sequence length (Figure 5.5). By defining only one partition that covers all target sequences a fixed value of n_p can be used, whereas classifying target sequences into multiple partitions allows a more fine-tuned and length-dependent assignment of n_p.

Table 5.1: *Comparison of three different design strategies to handle overlapping sequences. The table gives a crude estimation for various criteria, based on "typical" datasets with a medium level of overlapping target sequences. Trivially, for datasets with no overlaps among target sequences, all three strategies yield identical results. Probe-sequence coverage denotes the percentage of target sequences that are represented by at least one probe. See Figure 5.4 and text for details (in particular, what the names of the different strategies correspond to).*

Criteria	Strategy 1	Strategy 2	Strategy 3
Amount of overlapping probes	zero–high	zero	zero
Probe specificity with respect to individual target sequences	low–high	low–high	high
Probe-sequence coverage	high	low–high	low–high
Typical amount of probes that will be designed	medium–large	medium	few–medium

Typically, larger n_p values may be used for longer sequences, and particularly long target sequences may optionally be split into shorter subsequences. To ensure that probes may also be designed in the vicinity of the split positions, a specific overlap can be defined that should be at least as large as the probe length (Figure 5.5).
Application of that strategy therefore has multiple advantages. It can be very flexibly adjusted to best meet the requirements of the particular experiment and dataset, it uses the available array capacity efficiently while simultaneously controlling measurement bias if reasonably set, and it may also be used for traditional length-independent n_p values.

The `CAD` pipeline also provides various statistics to estimate the maximum number of probes that may be designed for a particular set of target sequences and design strategies. This is useful to obtain a crude estimate of the used space on the CEM after probe design and to evaluate if additional target sequences can be included. If the estimated number of probes exceeds the array capacity, for example, the number of target sequences may be reduced by discarding particular datasets, adding negative set filters, excluding sequences on the complementary strand for selected datasets, or by reducing the value(s) for n_p.

5.2.1.5 Probe Design

The design of high-quality probes is essential for accurate expression measurements [372, 388, 389]. Probes should have high specificity (to what degree they bind to non-target sequences) and sensitivity (binding strength to its target sequence). Thus, probes should have a high intensity if its target sequence is contained in the RNA sample and a low intensity otherwise. Furthermore, they must be isothermal with respect to their hybridization temperature [377] because they are all subject to identical experimental conditions.

I provide selected and preprocessed target sequences in universally accepted file formats that can

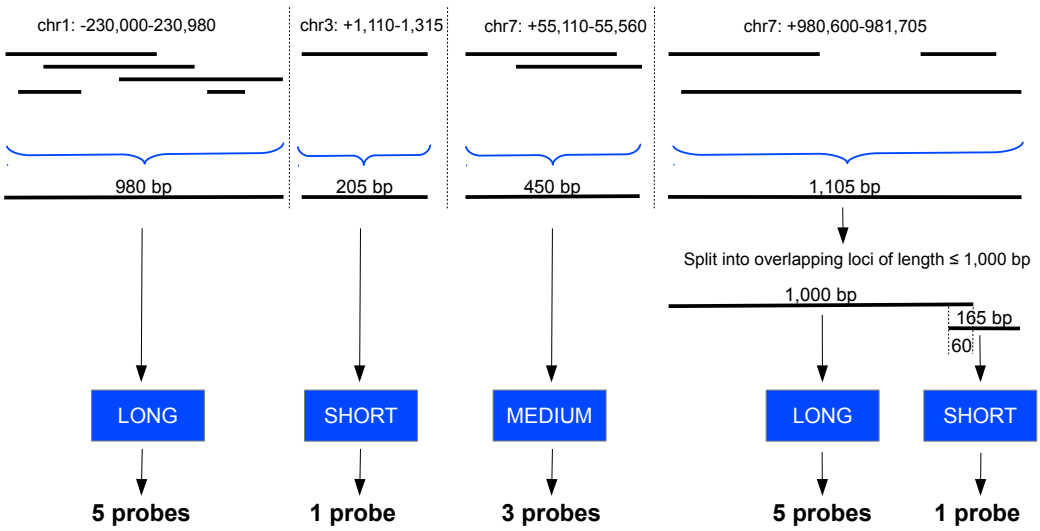


Figure 5.5: *Visualization of the first strategy to handle overlapping sequences in conjunction with target sequence partitioning into different sets. An exemplary set of (partly) overlapping target sequences is shown, originating from four different loci. Here, the first strategy is employed: overlapping target sequences are merged to one combined sequence and subsequently partitioned into three different sets (short: 60–300 bp, medium–long: 301–600 bp, and long: 601–1000 bp). If combined target sequences are longer than 1000 bp, they are split into subsequences of length $\leq$ 1000 with an overlap of 60 bp. The resulting target sequences may then be treated separately for subsequent probe design (e.g., in terms of the number of probes per target sequence). For more details, see text.*

be directly used as input for various probe design tools. For two reasons, I do not include the possibility to design oligonucleotide probes within the `CAD` pipeline. First and most importantly, CEM manufacturers like *Agilent* or *Nimblegen* provide their own models for probe design to optimize melting temperatures for their specific array technology. Second, although numerous software tools are available to design oligonucleotide probes (e.g., see [372, 389]), they are not suited to integrate with the `CAD` pipeline because they are:

- not freely available or require an official institutional agreement to obtain a free license (e.g., `Picky` [390], `Array Designer`[3])
- not available anymore or non-functional (e.g., `HiSpOD` [391], `EvoOligo` [392])
- cannot be downloaded and only run on a web server (e.g., `HiSpOD` [391], `ExonPrimer`[4])
- not available for Unix-based operating systems, cannot be run in a command-line mode, or are only accessible via a GUI (e.g., `Picky` [390], `Mprobe 2.0` [393], `PerlPrimer` [394])
- specialized for particular types of microarrays or application areas (e.g., `Teolenn` [395]: only

[3] `http://www.premierbiosoft.com/dnamicroarray`, last accessed in August 2013
[4] `http://ihg.gsf.de/ihg/ExonPrimer.html`, last accessed in August 2013

suitable for tiling arrays, ProDesign [396]: designed primarily for metagenomics)

- restricted with respect to the reference genome that may be used (almost all tools)

5.2.1.6 Probe Specificity

Cross-hybridization to similar sequences is typically the most relevant source of non-specificity. The threshold for when a probe is considered non-specific is non-trivial and depends on the probe length, for example. A good determinant for specificity is its similarity to other sequences [397], and cross-hybridization may already occur if the sequence similarity is as low as 75% or 70% (for probes of length 50 and 70, respectively) or if at least 15 contiguous bases are identical in sequence [397]. Therefore, evaluation of probe specificity is crucial, particularly because available probe designers only find probes that are unique with respect to currently annotated transcripts (e.g., *Agilent*'s *eArray*). Due to the pervasiveness of transcription (see Section 2.1.3), however, such a strategy is insufficient, and cross-hybridization must be tested against the complete genome to not overestimate probe specificity. Thus, although probes may map uniquely to RefSeq, for example, in reality, multiple perfect or near-perfect hits are frequently identified when mapped against the whole genome. This is caused by highly similar sequences not included in RefSeq that, however, may also be transcribed, making it impossible to pinpoint where the expression signal originally comes from.

In the CAD pipeline, probe specificity is therefore evaluated using the full genome as a reference. The mapping of probes to the genome is performed using BLAT [381], and its parameters may be arbitrarily customized to maximize flexibility. However, I encourage the usage of parameters that allow the splitting of probes (e.g., for probes overlapping a splice site in a particular RNA target sequence). Uniquely and non-uniquely mapping probes are provided separately, and non-uniquely mapping probes are discarded for further analyses. Although the removal of probes reduces the number of target sequences covered by at least one probe, it ensures that the origin of the expression signal is unambiguous.

5.2.1.7 Probe Coverage

After evaluating probe specificity and discarding non-specific probes, the analysis of probe coverage with respect to the target sequences is important. In particular, what is of interest is the percentage of target sequences that are represented by at least one specific probe (hereafter denoted $c_{>0}$). Generally, because each target sequence without a probe will not be represented on the CEM, $c_{>0}$ should be maximized and is therefore a measure for the probe coverage of the target sequences. Estimating $c_{>0}$ can principally be done based on different stages during the execution of the CAD pipeline: (i) after unification of all original target sequences but before any additional filters were applied, (ii) after the negative set filter were applied but before partitioning target sequences and

applying a strategy for handling overlapping target sequences, and (iii) after all steps of the CAD pipeline.

If $c_{>0}$ is not sufficiently high, additional probe designs with the same probe designer or runs of the CAD with altered parameters may be performed (e.g., increasing the number of probes per target sequence so that new potentially specific probes are produced and alternative design strategies, respectively). I discourage from using alternative probe designers, however, because they may design probes with very different chemical properties (such as hybridization temperatures).

5.2.2 The CEM-Designer Web Server

In addition to the CAD pipeline, a web server (CEM-Designer) was designed to make the CAD pipeline publicly available for other researchers (http://designpipeline.bioinf.uni-leipzig.de). It provides a user-friendly, flexible, and dynamic interface and is organized in three independent steps (Figure B.3): target sequences selection (step 1), and probe selection and design statistics (step 3). As discussed before, probe design (step 2) must be performed with external tools and is not integrated into the CEM-Designer. A major goal in the development of the web server was to design a user interface that minimizes manual user interactions and input but which can nevertheless be flexibly adjusted for more advanced users (Figure B.3 and Appendix B for details).

In step 1, an arbitrary number of files containing positions or sequences from genomic regions (target sequences) may be uploaded by the user. The user must then select and choose among various options (as discussed in Section 5.2.1) in the user interface of the CEM-Designer web server.
In step 3, probe specificity is evaluated (see Section 5.2.1.6), various statistics are calculated, and target sequences may be identified that are not yet represented by any uniquely mapping probes. Step 3 can optionally be performed independently of step 1, in which case only limited functionality is available (probe specificity evaluation only). Specifically, the following files and statistics are generated during the execution of step 3:

- **Probe-associated files and statistics**
  I provide various files that summarize all probes that could or could not be mapped uniquely (including a list of all found hits) or could not be mapped at all, both as BED and FASTA format. The files are in a human-readable, comma-separated text format and can therefore easily be opened with any spreadsheet program.

- **Target sequences-associated files and statistics**
  I provide various files that summarize the coverage statistics of all specific probes with respect to the target sequences at different stages during the execution of the CAD pipeline (see above). Furthermore, summary statistics are created that list target sequences that are represented by the envisioned number of probes, fewer probes, and no probes at all. I also generate graphical

representations of various statistics such as histograms (e.g., target sequence length and number of probes per target sequence), and the correlation of target sequence length and the number of overlapping probes.

To minimize waiting times, the CEM-Designer web server can handle several requests in parallel. It also implements a job system, allowing users to submit an arbitrary number of jobs that are started upon availability of computational resources (Figure B.3). If not enough resources are available for a newly submitted job, it will be placed in a queue. Jobs can furthermore be canceled at any time. Once a job is finished the user is notified by email. All result files that the CEM-Designer web server produced are accessible via a designated website and also available as a download.

The running time of individual jobs depends on the complexity of the input datasets and varies in step 1 and 3 from minutes to multiple days. However, jobs are aborted if they cannot be completed within 72 hours. The most time-consuming step is the mapping of probes and target sequences that are available only in FASTA format. This is particularly the case because mapping parameters may be arbitrarily user-adjusted, which can dramatically increase the running time as compared to the default mapping options.

5.2.3 The nONCOchip 2.0 and the Alzheimer Custom Array

Two CEMs have been designed using the CAD pipeline: the nONCOchip 2.0 and the Alzheimer Custom Array. The exact composition of both arrays, including methodological details for the application of the CAD pipeline, are described in Appendix B. Here, I only describe general application areas for the CEMs and highlight datasets of particular relevance.

Both the nONCOchip 2.0 and the Alzheimer Custom Array contain a comprehensive set of coding and particularly non-coding sequences. The nONCOchip 2.0 is especially designed for the detection of oncologically relevant ncRNAs. It contains known and predicted as well as experimentally detected ncRNAs that are regulated in the context of tumor-relevant signaling pathways such as *p53*, *STAT3*, and in the cell cycle. Due to the potential of ncRNA as biomarkers (see Section 2.1.4.3), the nONCOchip 2.0 may therefore be used as a cost-effective tool for their identification and development.

The Alzheimer Custom Array is in large parts identically composed as the nONCOchip 2.0 but differs in a few notable details. It is well suited for the identification of differentially expressed loci in AD (see Chapter 6) because it contains various disease-associated loci.

Notably, both arrays contain two ncRNAs datasets with chromatin-regulating functions.
The first dataset comes from Khalil et al. [201] and contains a set of lncRNAs originating from actively transcribed genes. They were identified based on their particular chromatin signature, namely the presence of promoter-associated H3K4me3 and transcription-associated H3K36me3

(K4-K36 domains). Using tiling microarrays at a resolution of 10 bp, the authors showed that in nearly 75% of all K4-K36 domains, multi-exonic ncRNA are present with four exons per domain on average. This adds up to a total of 4,860 exons[5] that were used as input for the `CAD` design pipeline. Khalil et al. [201] also showed that many lncRNAs expressed in *HeLa* cells are bound by chromatin-modifying enzyme complexes such as *PCR2* and *CoREST* ($\approx$ 20% and $\approx$13%, respectively) as well as *SMCX* (a histone H3K4me3 demethylase). Because *CoREST* is a repressor of neuronal genes, it is particularly interesting for the `Alzheimer Custom Array`.
The second chromatin-associated dataset comes from Mondal et al. [188] and contains 141 intronic and 74 intergenic regions from human fibroblast cells. The authors isolated, purified, and deep sequenced chromatin-associated RNA from fractionated chromatin and then focused on transcripts originating from intronic and intergenic transcripts. They subsequently successfully verified the chromatin-association property of a few of these putative transcripts and functionally examined one candidate further (see Section 2.1.4.3).

5.3 Discussion

Microarray technology is relatively mature and well-established. CEMs, in particular, are increasingly popular. However, mammalian transcriptomes are pervasively transcribed, with large numbers of overlapping transcripts for each traditional protein-coding gene (see Section 2.1.3). A suitable selection and preprocessing of target sequences prior to probe design as well as the evaluation of probe specificity are therefore non-trivial. On average, each traditional protein-coding gene has large numbers of overlapping transcripts (see Section 2.1.3), and protein-coding and non-coding transcripts often originate from the same genomic locus. Confidently measuring expression of particular non-coding variants is therefore challenging because it must be ensured that one can distinguish between non-coding and coding transcripts. Similarly, if multiple diverse datasets are integrated on the CEM, selecting a suitable strategy for how to treat overlapping target sequences and how to choose the number of probes per target sequence appropriately is a more complex task than previously appreciated. All of these considerations may have a profound impact on microarray data analysis and statistical validation and must therefore be addressed with great care.

The developed bioinformatics pipeline is the first tool that explicitly addresses these issues in a systematic and thorough fashion. It automates and facilitates the design of CEMs emphasizing in particular on target and probe selection. It provides high flexibility for the selection and preprocessing of target sequences because various specific design strategies with respect to handling overlapping target sequences or partitioning target sequences into smaller sets can be user-adjusted. It furthermore allows the identification of probes with high specificity for the desired target sequences and vice versa target sequences with low coverage of high-quality probes. To fully ensure high

[5] Table 2 in Dataset S1 in Khalil et al. [201]

probe specificity one would need the complete transcriptome, which is not yet available for complex transcriptomes. The approach presented here alleviates the shortcomings of available probe designers with respect to evaluating probe specificity by using genome-wide alignments that include alignments spanning introns. It also minimizes the effects of cross-hybridization to other target sequences, even those on yet unannotated transcripts that may nevertheless be contained in the RNA sample.

I introduced three different strategies to handle overlapping target sequences, each of which with its own application areas. Thus, their individual advantages and disadvantages have to be carefully considered. For the `nONCOchip` 2.0 and the `Alzheimer Custom Array` I found that a combination of these strategies provided a good compromise between representing a large fraction of target sequences with at least one probe and minimizing probe overlap. The large majority of the probes for both arrays originate from non-coding sequences. Importantly for the scope of this thesis, however, they also include genomic regions and transcripts with known or predicted chromatin-regulating functions.

As mentioned, to maximize the $c_{>0}$ value[6] I performed up to two additional designs after the first design using a reduced dataset containing only those target sequences that were not yet represented by any specific probe after the respective previous design. In summary, this scheme generally worked well. For Gencode lncRNAs, for example, the $c_{>0}$ value after the first design was ≈77%. Other classes of RNAs had substantially higher values (e.g., $c_{>0} \approx 96\%$ for coding exons and $c_{>0} \approx 88\%$ for UTRs). For the ncRNA datasets it was more difficult to design uniquely mapping probes. After the first design round, for example, the $c_{>0}$ value was only 66% (Figure B.4 A), and I improved it to ≈73% after the third design.

The overarching goal of the `CEM-Designer` web server is to provide a publicly available, easy to use, and flexible tool that implements the functionality of the `CAD` pipeline. User friendliness was particularly important, and the web server make use of various JavaScript libraries to optimize, for example, navigation and orientation, visual clarity, consistency, and comprehensibility. I expect that the web server itself will evolve over time to further increase flexibility and functionality. For example, it currently only supports `hg19` as target genome assembly version but additional genomes can easily be added if the need arises.

The `CAD` pipeline and the `CEM-Designer` web server may be further improved and extended by providing additional probe specificity tests such as more rigorous testing for potential cross-hybridization issues. For example, it has been shown repeatedly that cross-hybridization may occur if probes are too similar (>90%), identical in more than 15–20 consecutive positions [397, 398], or if the binding free energy is too low [398].

Despite the emergence of more advanced high-throughput methods such as RNA-seq, microarrays still offer multiple advantages. First, they are still more cost-effective compared with sequencing

[6] fraction of target sequences that are represented by at least one uniquely mapping probe

technologies although sequencing costs will continue to decrease. Second, microarrays are a matured technology. In contrast to RNA-Seq, sources of bias are better known, and powerful analytical strategies and experimental designs to deal with them are available [57]. Third, microarray data are easier and faster to analyze compared to RNA-Seq due to the immensely decreased size of the data. This reduction in complexity is particularly relevant for biologists and labs without extensive bioinformatics support. Fourth, microarrays have been found to perform better in various specialized tasks such as network reconstruction [399]. Thus, they will remain a useful, accurate, and ubiquitously used[7] tool for transcriptome profiling for at least a few more years.

[7] e.g., a *PubMed* search reveals that in 2012, the number of citations for microarray-related studies is over ten times higher than for RNA-Seq studies (search terms: "microarray" and "RNA-Seq" (or "RNASeq"), respectively)

Chapter 6

The Significance of the Chromatin Computer in Alzheimer's Disease

6.1 Motivation and Background

As outlined in previous chapters, the maintenance of correct epigenetic patterns throughout the lifetime of an organism is crucial for cellular stability and identity. Misregulation of epigenetic mechanisms is likely to have fatal consequences that may contribute significantly to diseases such as Alzheimer's disease (AD). Indeed, epigenetic mechanisms have been increasingly associated with AD. Consequently, they are of growing importance in the etiology of AD [42, 43, 48–54]. Zawia et al. [48], for example, suggested that the promoters of AD-associated genes may be hypomethylated due to environmental influences during brain development that cause the inhibition of DNA-methyltransferases, followed by a cascade of events such as increased rate of DNA damage and neurodegenerative processes. Lithner et al. [400] hypothesized that soluble amyloid-ß may be a signaling molecule that modulates the transcriptional activity of DNA by disrupting histone H3 homeostasis. Lastly, Rao et al. [50] measured global AD-associated DNA hypermethylation and histone H3 phosphorylation in genomic regions that involve, among other functions, neuroinflammation.

In summary, DNA methylation and histone PTMs are promising candidates for further investigation and drug development [48–51], as well as long non-coding RNAs and particularly chromatin-associated RNAs [52–54]. Additionally, histone acetylation is also believed to contribute to the pathogenesis of AD and has recently spurred substantial interest due to its importance for learning and memory [401]. Hypoacetylation is generally associated with repressive chromatin states, and hypoacetylation of histones H3 and H4 in particular may initiate apoptosis in neurons [402]. Histone deacetylase inhibitors in particular are therefore currently promising candidates for treating central nervous system disorders [403, 404].

A multitude of putative causes for AD have been postulated. Various genetic studies suggest that AD is not caused by a simple mutation, because only a very small percentage of AD cases can be linked to mutations in specific genes. Instead, it has been hypothesized that the cognitive decline is

caused by abnormal network-related activities that interfere with cognitive functions such as learning and memory, apoptosis of particular neuronal populations, as well as dysfunction and aberrant loss of synapses [405]. More specifically, a plethora of gene expression profiling studies revealed dysregulation of genes related to neuroinflammation, intracellular signaling pathways such as calcium, zinc or Wnt signaling and signal transduction [406, 407], metal ion binding and dyshomeostasis [408, 409], mitochondrial dysfunction [410], cell cycle regulation [411, 412], membrane integrity [413], the immune system and inflammation [414], protein kinases and phosphatases [415], neuronal and synaptic plasticity as well as neurotransmission [416], the cytoskeleton [417], cell adhesion [418], and olfactory dysfunction [419]. However, some of the observed dysregulations may only be downstream effects (e.g., neuroinflammation [410]). In addition, the pathological cascade of AD seem to begin at least ten years before the appearance of the first clinical symptoms [420–422].

For AD and age-related neurodegenerative disorders more generally, the prion hypothesis has recently gained experimental momentum [422]. Briefly, prions are typically infectious agents composed of misfolded proteins. These so-called proteinaceous seeds may serve as self-propagating agents for the instigation and progression of particular diseases. In AD, ultimately, they functionally compromise the nervous system due to aggregated proteins that either gain a toxic function and/or lose their normal function [422]. Whereas prion diseases in general may also be infectious in origin, this does not seem to be the case for AD.

Biochemically, AD is associated with amyloid plaques and neurofibrillary tangles in the brain [423] but a clear causative pathway is yet missing [47] (Figure 6.1). As an example for the causal complexity, Ciarlo et al. [424] demonstrated that the ncRNA *51a*, located antisense to an intron of the AD risk gene sortilin-related receptor 1 (SORL1), is frequently upregulated in AD. This promotes the synthesis of an alternative splicing variant of SORL1, which subsequently causes increased ß-amyloid formation [424].

AD seems to be an evolutionarily young disease and is currently believed to occur only in humans. No other mammalian species recapitulates all of the key features of AD [55] although some are also susceptible to AD-like symptoms [55, 56]. Data availability for non-human primate species is very rare, and it therefore remains unclear to what extent AD is indeed human-specific. An intriguing hypothesis concerns whether the genomic loci that are differentially regulated in AD are similarly evolutionarily young (hypothesis 1). To this extent, I am interested in identifying whether they show signs of recent changes in their genomic structure. It is reasonable to speculate that this is particularly true for non-coding regions and only to a much lesser extent for protein-coding ones. Because the latter have high evolutionary pressure, they are largely conserved across mammals and evolutionarily much more conserved.

Additionally, if epigenetics indeed plays decisive roles in AD as suggested by the large array of studies, chromatin-associated loci and loci associated with epigenetic stability should be differentially expressed in AD. I therefore hypothesize that a dysregulated CC may play important roles in AD

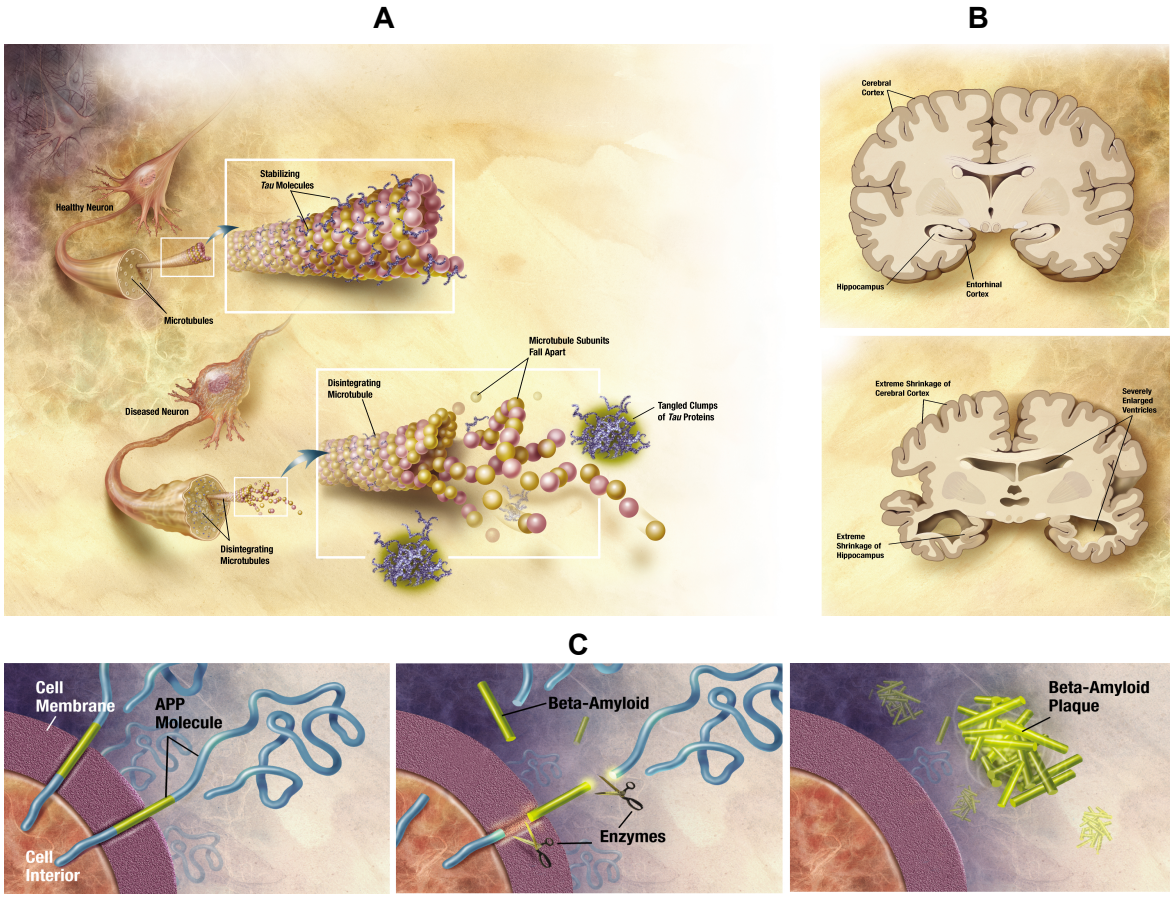


Figure 6.1: *Pathophysiology of AD.*
A: Schematics of how microtubules disintegrate in AD brain cells caused by changes in the tau protein, which destroys the neuron's transport system [425]. Microtubules are essential for the cytoskeleton of every neuron. Credit: National Institute on Aging/National Institutes of Health.
B: Comparison of two brain diagrams (top: normal brain, bottom: AD brain). Credit: Image is released to public domain.
C: In AD, amyloid precursor protein, a single-transmembrane protein present in neuronal synapses with important neuron-related functions, is cleaved into fragments. Most notably, this yields beta-amyloid, which forms amyloid plaques (insoluble fibrils that accumulate outside neurons) that are characteristic for AD. Credit: NIH National Institute on Aging.

(hypothesis 2).

In this chapter, I explicitly address the two hypotheses about the evolution of AD and its association with chromatin using the `Alzheimer Custom Array` that has been developed (see Section 5.2.3). Due to the complexity of the bioinformatic analyses that I performed, for some analyses, full methodological details are listed in Appendix C and only a brief summary of the methods is given in the main text.

6.2 Methods and Results

6.2.1 Microarray Workflow

Analyzing microarray data poses significant bioinformatics challenges, for example in terms of the large amount of data that has to be analyzed, the assessment of data quality and comparability, and finally in the obtainment of accurate and reproducible results. Figure 6.2 depicts a general schematic overview of a typical microarray workflow, and as highlighted, only steps 4–8 are described in the following.

Steps 1–3, particularly the preparation and microarray processing of the control and AD samples for the `Alzheimer Custom Array`, were performed in the RNomics group at the *Fraunhofer Institute for Cell Therapy and Immunology* (Fraunhofer IZI, Leipzig, Germany) in collaboration with the *Paul-Flechsig-Institute for Brain Research*, Leipzig, Germany. Methodological and experimental details are therefore not described here but can be requested from the following individuals:

- Dr. Kristin Reiche (`kristin.reiche@izi.fraunhofer.de`)
- Prof. Dr. Thomas Arendt (`Thomas.Arendt@medizin.uni-leipzig.de`)
- Dr. Jörg Hackermüller (`joerg.hackermueller@izi.fraunhofer.de`)

Steps 4–8 were performed using the statistical software package `R` and `Bioconductor` [426]. After the identification of differentially expressed probes, I employed a variety of additional methods and analyses to unravel the significance and meaning of the expression results and to relate targets to biological functions.

6.2.2 Quality Control and Data Normalization

Quality control measures encompass a variety of methods and techniques to critically assess the quality of each array in order to identify artifacts of any kind. Upon visual inspection of the arrays and based on sample similarity heatmaps, I excluded four arrays because they displayed spatial artifacts, thereby improving data quality. For example, these arrays contained regions with artifactually low or high intensities relative to the majority of the array [427]. Using sample similarity heatmaps, I furthermore uncovered that the age of the patient contributed non-negligibly to array similarity and subsequently included the variable age in the linear model for the identification of differentially expressed probes (see Section 6.2.3). For noise reduction (i.e., removing probes with consistently low intensity values or low variance across all samples), I filtered probes if their intensity scores had (i) an interquartile range[1] of less than 1 (unspecific filtering) or (ii) an expression

[1] a common statistical measure that is equal to the difference between the upper and lower quartiles of the intensity scores for each probe (i.e., a measure of the spread of the middle 50%)

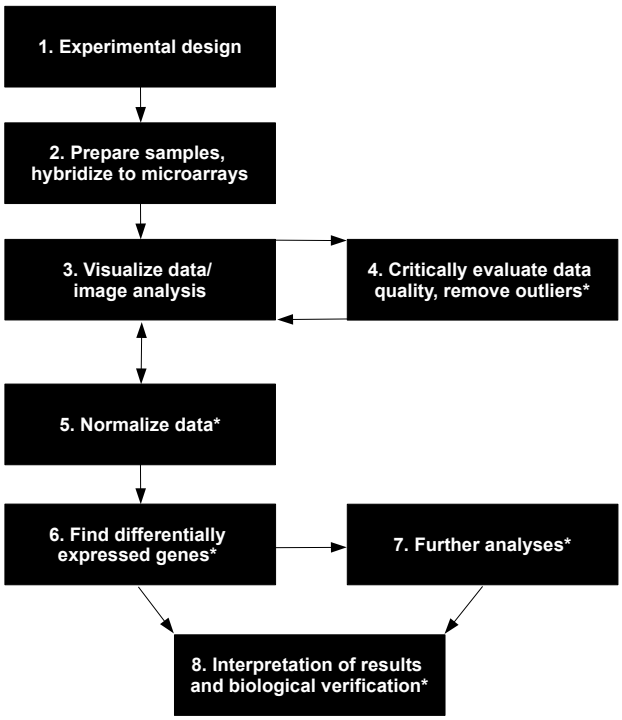


Figure 6.2: *Schematic of tasks and steps in a classical gene expression microarray experiment. Only steps marked with an asterisk are described here. Image modified after Slonim et al. [378].*

determination above the background in less than one third of all arrays. From the original 931,898 probes, this retained 113,047 that passed all quality filters.

Although performing a background correction is useful because it eliminates unspecific background noise and local fluctuations of the overall signal level, it typically also lowers signal intensities. Because Agilent microarrays already have relatively low signal intensities, I therefore decided against a background correction.

Data normalization is crucial to reduce systematic sources of variation (technical variation) such as among-array differences in the amount of starting RNA and the efficiency of various experimental issues (photodetection, reverse transcription), and other systematic biases that hinder or prevent meaningful biological comparisons. Also, comparability among experiments and arrays within one experiment is increased. Importantly, biological variation should be left untouched. I decided to quantile normalize expression intensities [428].

6.2.3 Identification of Differentially Expressed Probes

To identify differentially expressed probes between AD and control samples, I defined a linear model that explicitly includes the patient age because on average, individuals from AD samples were much older as compared to the control samples (∼81 and 65 years, respectively). The model also included error terms to account for stochastic variation. For more details, see Appendix C.1. I then identified differentially expressed probes based on different q-value thresholds to control the false discovery rate (0.2: 4,184; 0.1: 2,021; 0.05: 1,038; 0.01: 214). In Figure 6.3, I show two example heatmaps based on $q = 0.1$ and $q = 0.2$. Finally, I chose a q-value threshold of 0.2. I set q to a relatively high value because I employed a two-step procedure to identify differentially expressed loci (see Section 6.2.5). Of the 4,184 differentially expressed probes with $q < 0.2$, 3,263 were upregulated in AD and 4,095 mapped uniquely. The latter defined the set s_{diff} that I used for all subsequent analyses.

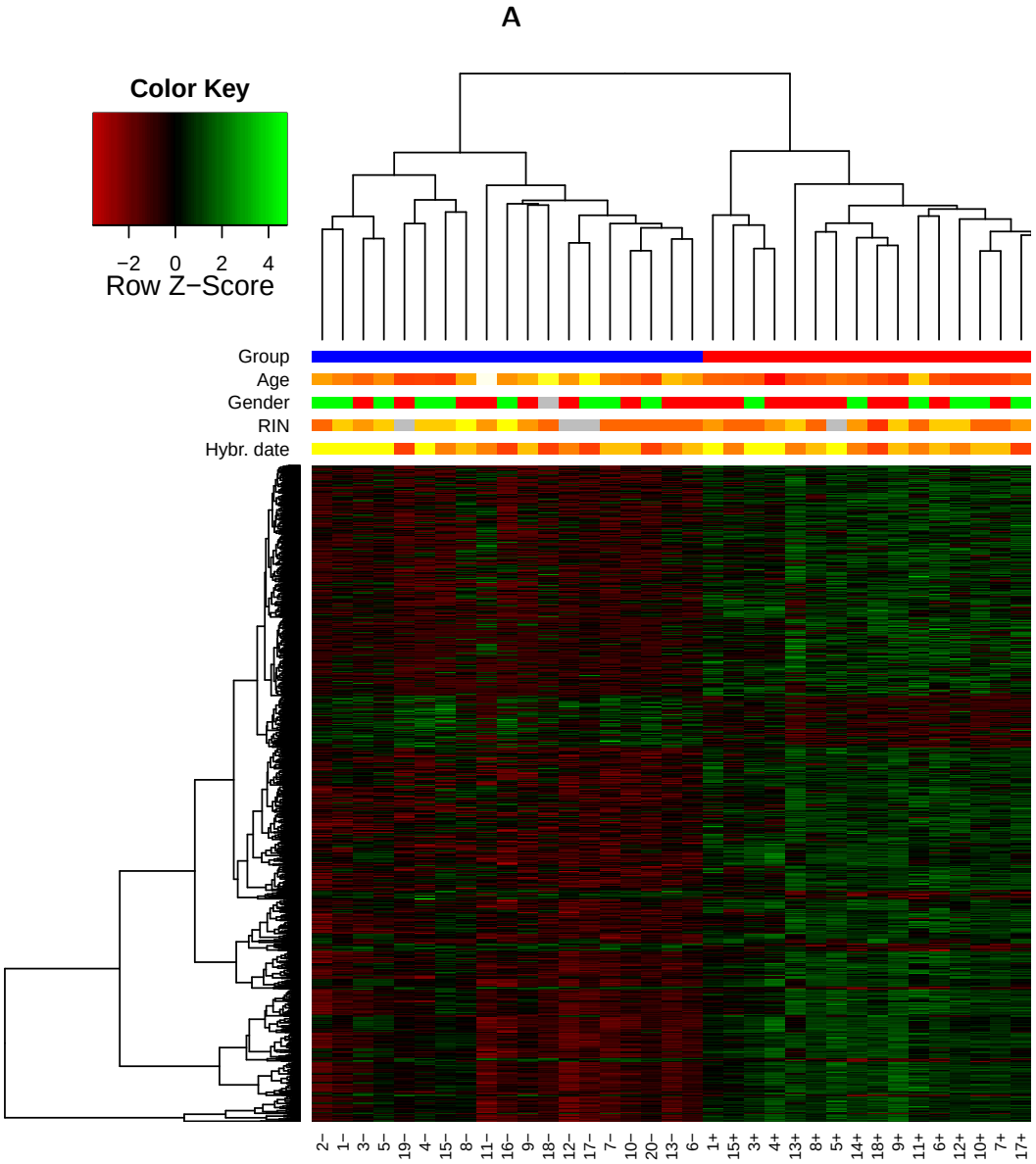
6.2.4 Feature Enrichment Analyses

I then performed a feature enrichment analysis for s_{diff} (e.g., analyzing the location of probes with respect to specific genomic features such as exons, introns, and UTRs).

I performed the enrichment analysis based on the *Gencode* v14 annotation for the following three annotation fields: feature type, gene type and transcript type[2] and the various annotation (source) files that I integrated (see Appendix B.2 for details). In addition to the feature types as provided by Gencode, I also introduced three additional ones (intron, intergenic and lncRNA). Introns were defined as intragenic regions that are not exons. Intron positions were derived from Gencode, based on the positions of exons and genes. Intergenic regions were defined as regions that overlap with neither any protein-coding transcript nor any kind of pseudogene. To estimate the number of probes that do not overlap with any protein-coding transcript but with a pseudogene, I additionally included a second intergenic category that ignores any pseudogene overlap and therefore only incorporates the protein-coding overlap information (labeled as *Intergenic (ignoring pseudogenes)* in Figure 6.4). lncRNA finally indicates that the probe overlapped with a transcript from the *Gencode* long non-coding RNA subset. Except for intergenic regions, I distinguished between sense and antisense hits for all overlaps. For the calculation of the observed frequencies, I counted the number of (distinct) probes that overlap at least 95% of their length (i.e., 57 nucleotides) with a particular annotation feature or dataset (see Appendix B.2). Because categories were not mutually exclusive, individual probes might be associated with multiple categories.

The results of the enrichment analysis based on the Gencode fields feature type, gene type, and

[2]The feature type can have the following values: gene, transcript, exon, CDS, UTR, start codon, stop codon, or Selenocysteine. For valid gene and transcript biotypes, see `http://www.gencodegenes.org/gencode_biotypes.html`, last accessed in June 2013



(Continued on next page.)

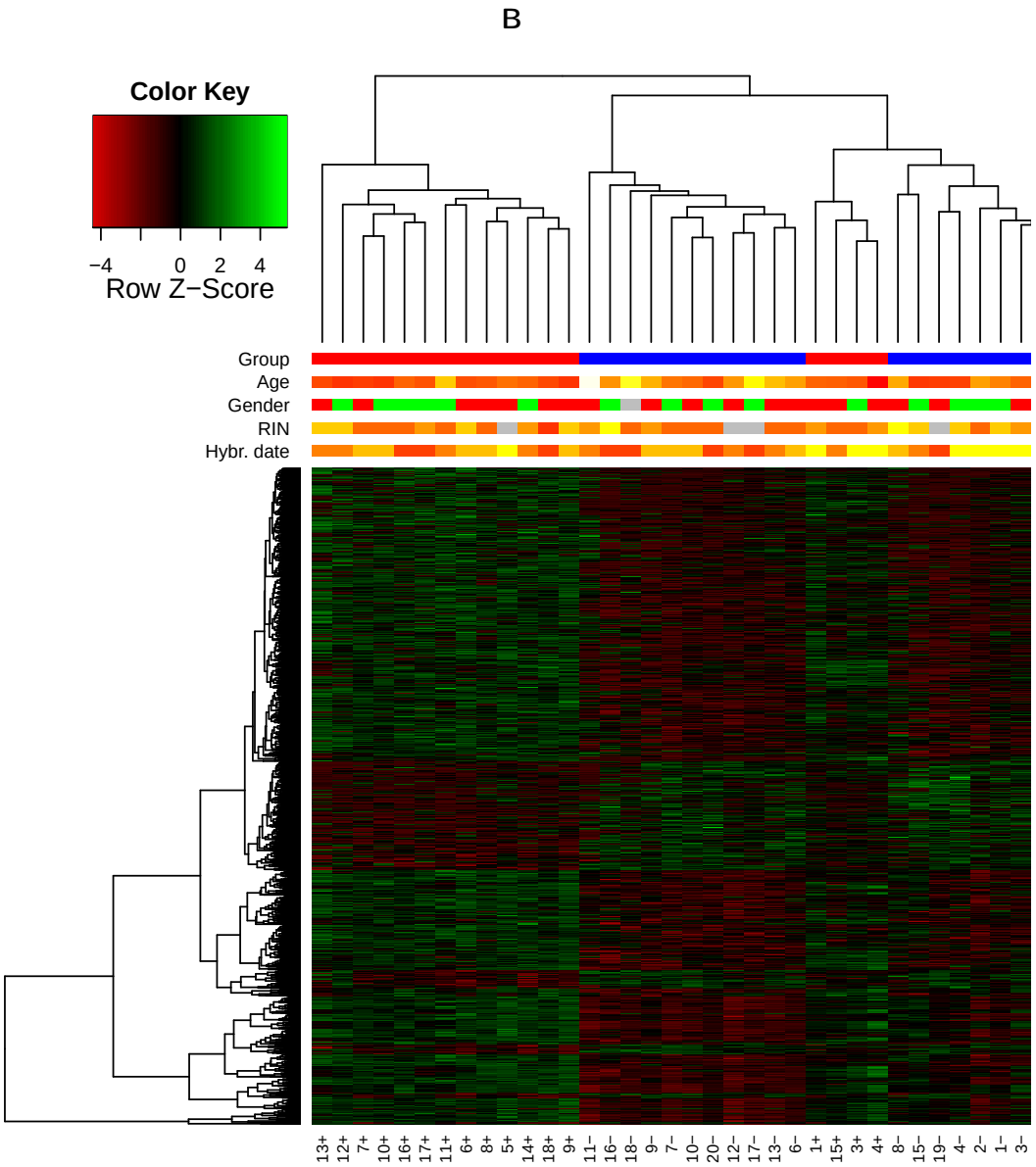


Figure 6.3: *Example heatmaps of differentially expressed probes. Two heatmaps (with dendrograms) of differentially expressed probes are shown, based on normalized expression values for $q = 0.1$ (A) and $q = 0.2$ (B). Each row shows one differentially expressed probe and each column a particular array (i.e., patient). For each array, additional metadata, such as patient group (blue: AD, red: control), age, and gender (green:male, red:female), are depicted as well as experimental metadata associated with the array of the patient such as RIN value and array hybridization date. The original values for age, RIN value and array hybridization date (AHD) were subject to a linear transform and are colored on a yellow-red scale (yellow/bright: young/low RIN/early AHD, red/dark:old/high RIN/late AHD, gray indicates missing data). Variable ranges: age:21–98 years, RIN:5–8.3, AHD:one of four possible consecutive days. For visualization purposes, values have been centered and scaled in the row direction (Figure C.1 for a variant without scaling). See text for further details. (Continued from previous page.)*

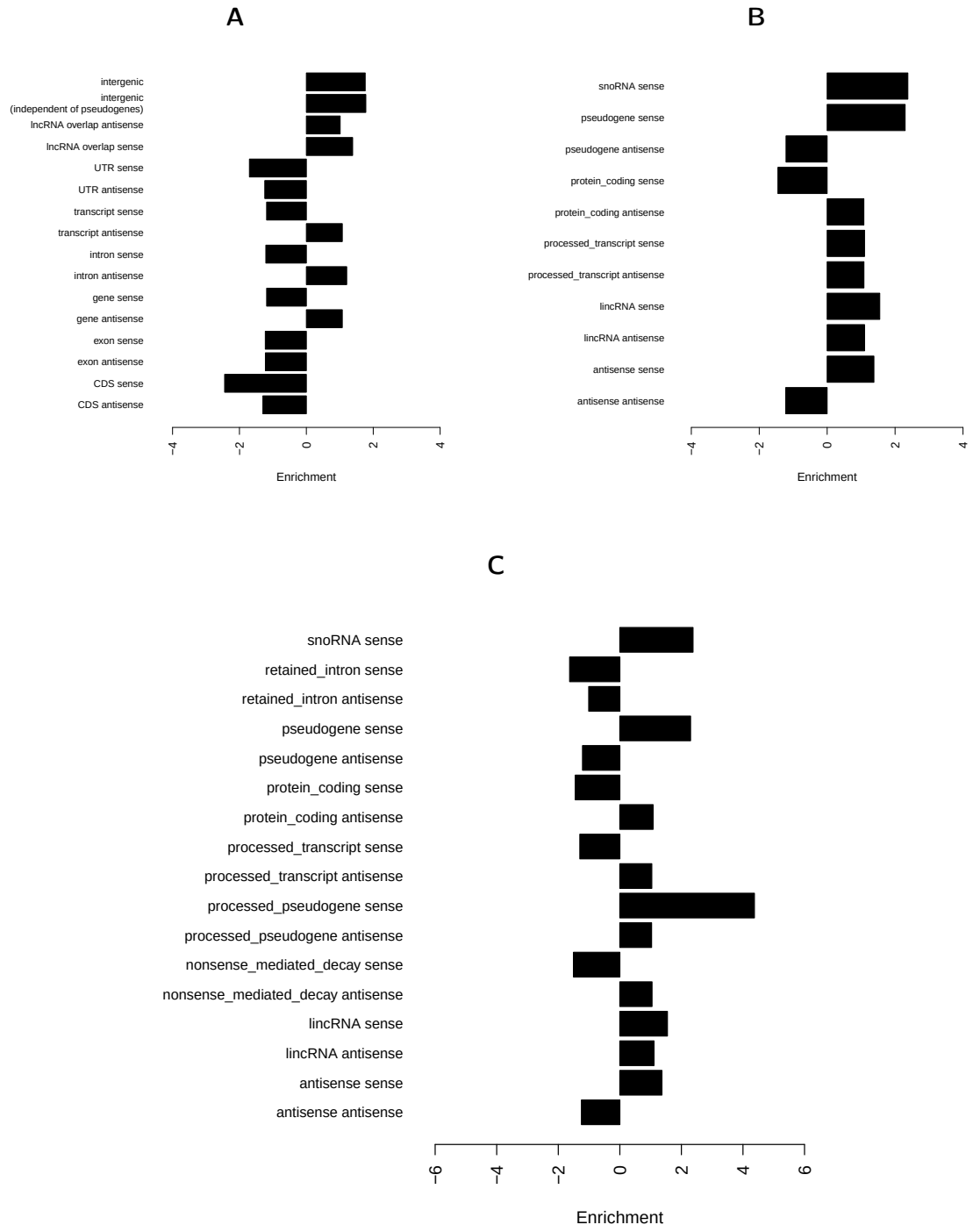
transcript type are summarized in Figure 6.4 A–C. For the feature type, I found an enrichment of intergenic regions, lncRNAs (sense), and introns (antisense). I also found less hits than expected for UTRs, coding regions and exons (sense and antisense) and introns (sense). For almost all categories, I found less sense hits than expected, while antisense transcripts were mostly enriched. The gene type enrichment analyses revealed an increased occurrence of a diverse set of ncRNAs such as snoRNAs, lncRNAs, and antisense hits of protein-coding regions. I found that pseudogenes and particularly processed pseudogenes were also highly enriched.

The results of the enrichment analysis based on the source files the regions originated from (see Appendix B.2) are summarized in Figure 6.4 D and E. I found enrichment of TP53 (sense and antisense), snoRNAs, ncRNA predictions based on `RNAz` and `Evofold`, and totally and partially intronic RNAs as identified by Nakaya et al. [429] (denoted as TINs and PINs, respectively). Intriguingly, both chromatin-associated datasets and the vast majority of the cell cycle tiling array regions (see Appendix B.2) were also enriched both in sense and antisense direction, with few exceptions.

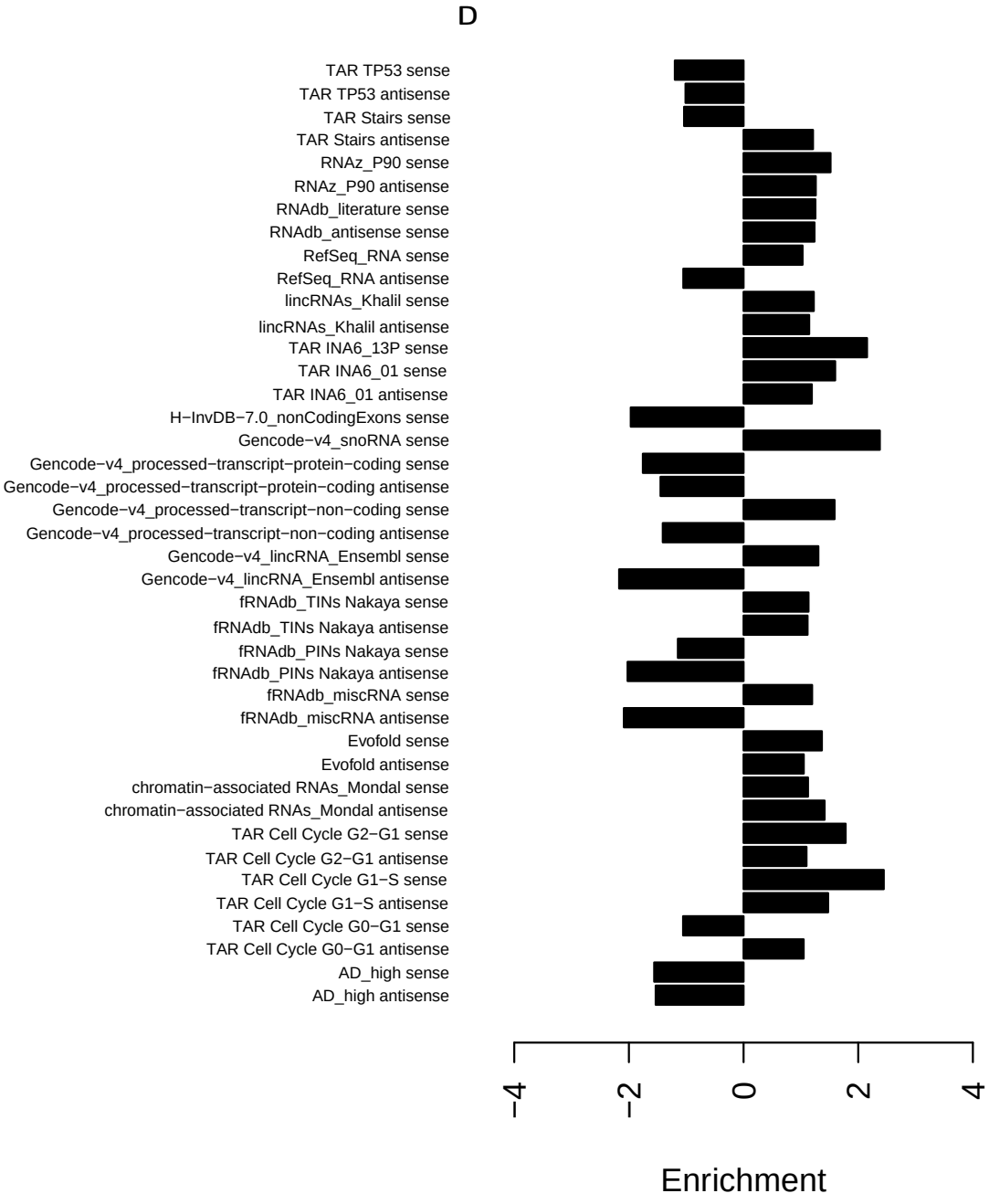
6.2.5 Identification of Differentially Expressed Loci

Next, I aimed to identify differentially expressed loci rather than probes. Due to the complexity of the methods, I only provide a brief methodological summary here, and the full details are described in Appendix C.2. The rationale behind this step was to identify the set of genes that show particularly trustworthy signs of differential expression. I argue that the differential expression of an individual probe may not be a sufficient criterion for the corresponding gene to be considered differentially expressed. For example, consider the following case for a particular gene g for which one differentially expressed probe p_{diff} mapping to g has been identified. Among all probes that map to g, p_{diff} may be a false positive, and all other probes do not show signs of differential expression. Thus, further incorporating g may not be useful because other genes show much stronger and homogeneous signals with respect to differential expressions of probes.

Briefly, I employed a two-step procedure for the identification of differentially expressed loci. The first step encompassed the identification of differentially expressed probes as described above, based on a relatively high q-value threshold. The second step entailed rigorous testing of all loci for which at least one probe $p \in s_{\text{diff}}$ maps in sense direction. Only loci were classified as differentially expressed if a significant proportion (based on a p-value threshold of 0.05 using a binomial test) of probes mapping to that loci had an expression change in the same direction as $p \in s_{\text{diff}}$. Thus, the false discovery rate was also controlled in a two-step fashion.



(Continued on next page.)



(Continued on next page.)

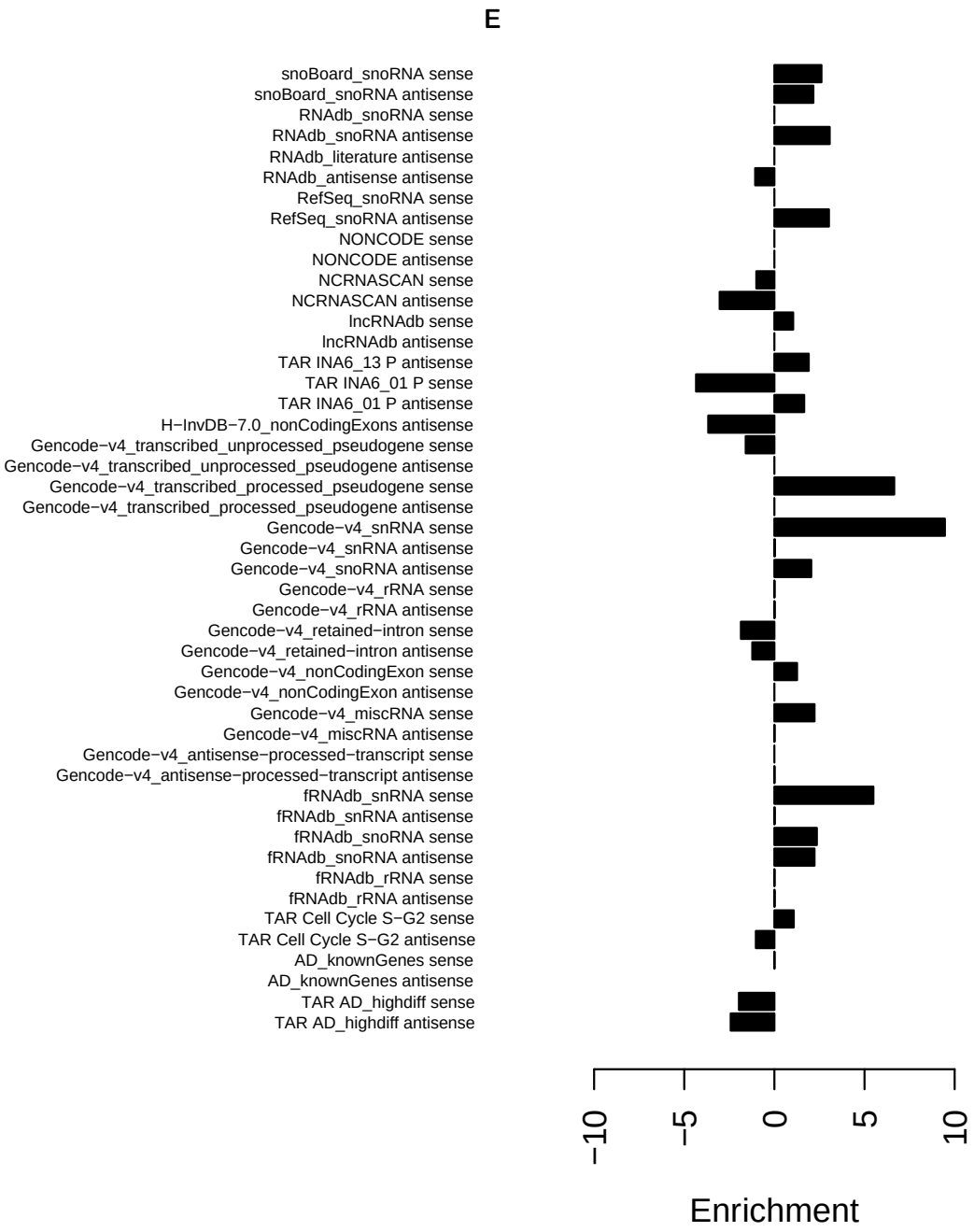


Figure 6.4: *Results of the enrichment analysis.*
A–C: Expected and observed frequencies with respect to the Gencode *annotation for the three* Gencode *annotation fields feature type (A), gene type (B), and transcript type (C) are shown (separately for sense and antisense overlap). Only types for which either the background or the observed frequency were larger than 10 are shown.*
D–E: Expected and observed frequencies with respect to the manually collected datasets (separately for sense and antisense overlap). In D, only datasets for which either the background or the observed frequency were larger than 10 are shown, whereas in E, the remaining datasets are depicted.
(Continued from previous page.)

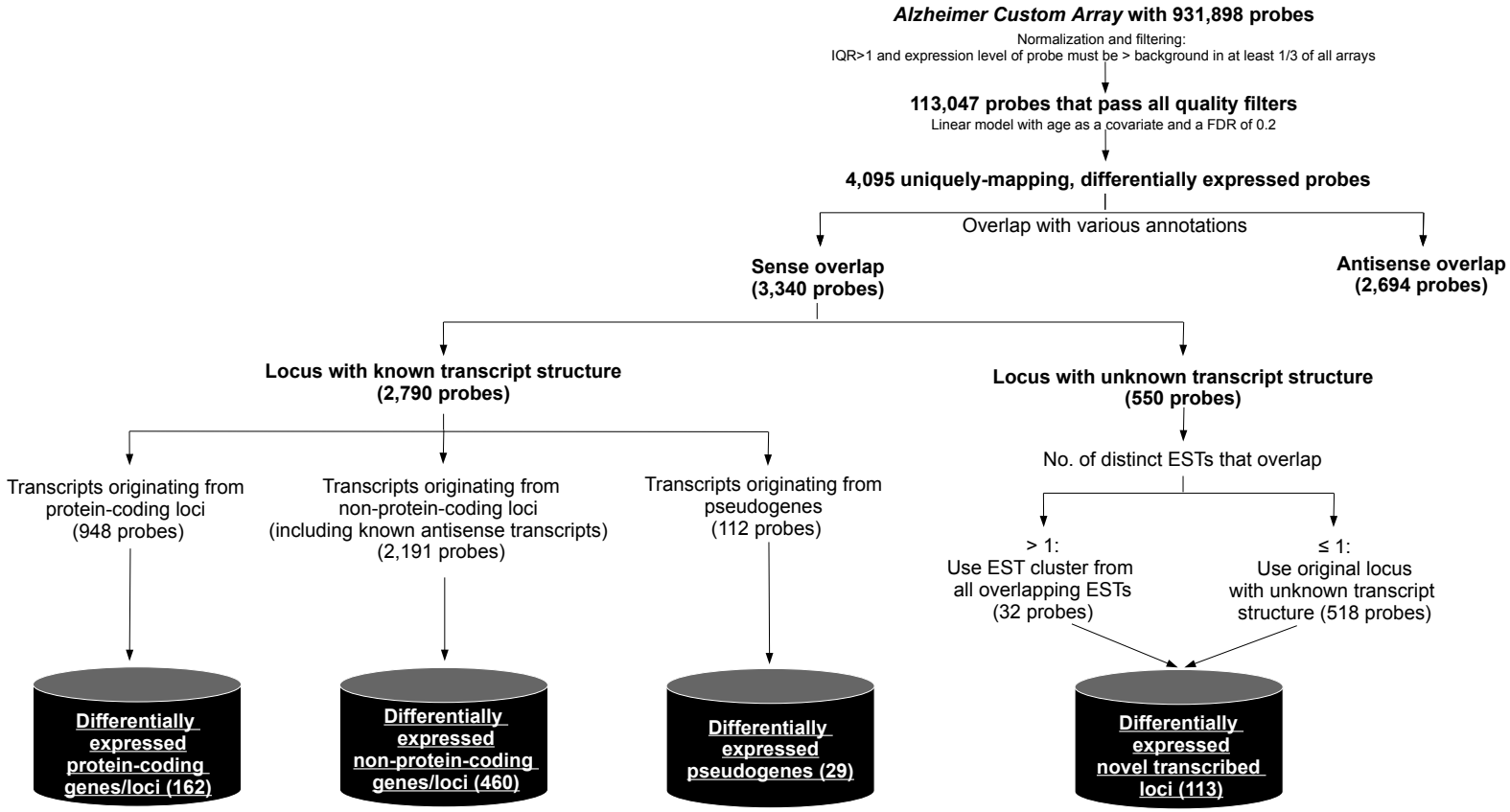


Figure 6.5: *Summary of the results and the methodology for identifying differentially expressed loci. For details, see text.*

Collectively, I identified a set of 764 differentially expressed genomic loci, 31 of which were associated with at least three distinct differentially expressed probes (Table 6.1). I classified each genomic loci to one of the four following classes: protein-coding, non-coding, pseudogenes, and uncharacterized. The first three classes correspond to known transcripts, whereas the latter represents loci with uncharacterized transcript structure and strand. They were previously identified by transcriptome-wide expression variation studies utilizing Affymetrix Human Tiling Arrays (tiling array regions — TARs, see Appendix B.2), RNAs with conserved secondary structures, and caRNAs as identified by Mondal et al. [188]. For details how I assigned a class for each loci, see Appendix C.2. In summary, the 764 loci split up in the four classes as follows (Figure 6.5):

- 162 putative protein-coding genes
- 460 putative non-coding genes or non-coding loci
- 29 putative pseudogenes
- 113 loci with unknown/uncharacterized transcript structure and type

Initially, a total of 208 loci were found that originated from annotations with unknown transcript structures and strand (loci from the tiling array experiments, ncRNA predictions, caRNAs [188]). However, only 113 of them did not overlap with any other annotation with known transcript structure (e.g., *Gencode* or any of the non-coding databases that I included, see above), whereas 95 overlapped with known transcripts. The former therefore require further study. For the latter, I used the overlapping annotation with known transcript structure and strand and ignored the overlap with the annotations with unknown transcript structure. Of these 113 (208) loci, 19(25) correspond to tiling array regions with putative cell cycle regulating functions (G0-G1: 11 (17), G1-S:1, G2-G1:7), 25 (33) to tiling array regions that were significantly higher expressed in a preliminary study with one AD patient sample and one control brain tissue sample, 18 (24) tiling array regions with RNAs that have been found to be controlled by major cancer-related pathways, 7 (10) caRNAs as identified by Mondal et al. [188], 42 (48) RNAs with conserved secondary structures (39 (45) `RNAz` [384], 3 `Evofold` predictions [382]), and 10 (89) operational genes (EST cluster).

6.2.6 Functional Characterization of Differentially Expressed Loci and Overlap with Known AD-Associated Loci

For differentially expressed protein-coding and non-coding genes, I collected functional descriptions of each gene using the `GeneCards`[3] [430, 431] and *Ensembl*[4] website. Additionally, I used the Gene Ontology (`GO`) project [432] to retrieve `GO` terms for each gene.

I found that most of the differentially expressed loci have known AD associations or hypothesized

[3] `http://www.genecards.org`, last accessed in August 2013
[4] `http://www.ensembl.org`, last accessed in August 2013

Table 6.1: *Selected differentially expressed loci. Differentially expressed loci associated with at least three distinct differentially expressed probes are listed.* ***Gene name****:* Ensembl *gene name (if available).* ***ID/Location****:* Ensembl *ID (if available) or chromosomal location.* ***Signal origin****: Characterization of the location/source from which the signal (differentially expressed probe) originates from (TAR — tiling array region, see Appendix B.2).* ***Reg. in AD****: Is the transcript downregulated (-) or upregulated (+) in AD in the* `Alzheimer Custom Array`*?* n_{diff}*: Number of distinct differentially expressed probes associated with the locus.*

Gene name	ID/Location	Signal origin	Reg. in AD	n_{diff}
SLC1A2	ENSG00000110436.6	exon of protein-coding gene	−	15
INHBA-AS1	ENSG00000224116	antisense lncRNA	−	12
RAB3B	ENSG00000169213.5	exon of protein-coding gene	−	11
C1orf95	ENSG00000203685.5	exon of protein-coding gene	−	10
NA	chr11:+35272757–35277591	uncharacterized (AD-associated TAR)	−	10
FREM2	ENSG00000150893.9	exon of protein-coding gene	−	9
NA	chr11:+35277781–35280190	uncharacterized (AD-associated TAR)	−	9
SDC4	ENSG00000124145.5	exon of protein-coding gene	−	7
AGXT2L1	ENSG00000164089.4	exon of protein-coding gene	−	6
NKAIN3	ENSG00000185942.7	intron of protein-coding gene	−	6
CTD-2015H6.1	ENSG00000240003.2	pseudogene	+	6
NA	chr10:+129901418–129907741	uncharacterized (cell cycle- and cancer-associated TAR)	+	6
NA	chr18:+41591002–42006002	uncharacterized (cancer-associated TAR)	+	6
PDGFRA	ENSG00000134853.7	exon of protein-coding gene	−	5
TEA	chr14:+22942568–23016566	lncRNAdb (ncRNA from the T early alpha promoter)	+	5
MIR31HG	ENSG00000171889.3	lncRNA	+	5
ZCCHC12	ENSG00000174460.3	exon of protein-coding gene	−	4
ZIC4	ENSG00000174963.13	exon of protein-coding gene	+	4
TARBP1	ENSG00000059588.5	exon of protein-coding gene	−	4
RUNX1	ENSG00000159216.13	intron of protein-coding gene	+	4
ZIC4	ENSG00000174963.13	exon of protein-coding gene	+	4
NA	chr11:+35280523–35282998	uncharacterized (AD-associated TAR)	−	4
NA	chr3:+123380693–123498543	uncharacterized (caRNA [188])	+	4
TTC6	ENSG00000139865.12	exon of protein-coding gene	−	3
ADAM12	ENSG00000148848.9	intron of protein-coding gene	+	3
FLI1	ENSG00000151702.11	intron of protein-coding gene	+	3
FRMD6	ENSG00000139926.11	intron of protein-coding gene	−	3
KIAA1199	ENSG00000103888.11	intron of protein-coding gene	+	3
CRIM1	ENSG00000150938.5	intron of protein-coding gene	−	3
COL6A3	ENSG00000163359.10	intron of protein-coding gene	+	3
LL22NC03-13G6.2	ENSG00000224404.1	lncRNA	−	3
EXT1	ENSG00000182197.6	intron of protein-coding gene	−	3

implications in AD neuropathology. Examples include *CALHM1* [433], *BDNF* [434, 435], *SLC1A2* [436], *KCNIP4* [52], *RAB3B* [437], *TP53* [438], and various other genes related to different signaling pathways such as calcium, zinc or Wnt signaling and signal transduction; metal ion binding and dyshomeostasis; cell cycle regulation; membrane integrity; the immune system and inflammation; protein kinases and phosphatases; neuronal and synaptic plasticity; neurotransmission; the cytoskeleton; cell adhesion; and olfactory dysfunction (see Section 6.1).

I found differentially expressed probes that map either in sense or antisense direction to various other previously reported AD-associated genes such as *FYN* (ENSG00000010810, antisense) [439], *PICALM* (ENSG00000073921, sense) [440], *NOVA1* (ENSG00000139910, antisense) [114], *GABBR2* (ENSG00000136928, sense and antisense) [441], and *LRP6* (ENSG00000070018, antisense)[442]. Particularly noteworthy are the following genes, for which probes in sense direction were classified as differentially expressed:

- **CALHM1**: This gene encodes a transmembrane glycoprotein that regulates cytosolic Ca^{2+} concentrations and amyloid-ß levels. Particular polymorphisms have been found to be significantly associated with AD [433].
- **BDNF**: see Section 6.2.7 for details
- **SLC1A2**: This gene has crucial roles for synaptic activation and for neuronal damage prevention from excessive activation of glutamate receptors. Its observed downregulation in AD may result in a low conductance state of synapses, which could be indicative for the activation of inhibitory pathways [443].
- **KCNIP4**: This gene encodes a potassium channel-interacting protein and regulates neuronal excitability in response to intracellular calcium changes. Recently, a ncRNA located antisense to an intron of *KCNIP4* has been identified that drives the synthesis of an alternatively spliced form of the gene. Upregulation of this particular isoform results in various dysregulated biochemical outcomes that may contribute significantly to brain homeostasis and pathogenesis [52]. *KCNIP4* also interacts with presenilin, one of the few genes with known mutations that may cause AD [444].
- **TP53**: see Section 6.2.7 for details

I also analyzed overlap with known AD-associated loci more systematically using a total of 451 genomic loci (genes) that have been collected from the literature (see Appendix C.3 for details). I successfully designed probes for all these 451 loci, each of which is covered by a median number of 21 probes (range:0–1,106). For 444 of them, at least two probes have been designed. I then determined the set of AD-associated genes that overlap with the set of differentially expressed loci. In summary, I found 22 overlaps in sense and 9 in antisense direction, thereby confirming their AD-association (Table C.1).

I also identified several putative protein-coding genes that are so far either uncharacterized or of unknown function but with potential pathogenic significance in AD. This list includes, for example, *TTC6* (ENSG00000139865), *RP1-27O5.3* (ENSG00000215897), *RGSL1* (ENSG00000121446), *GCSAML* (ENSG00000169224), *CLLU1OS* (ENSG00000205057), *CCDC60* (ENSG00000183273), *KIAA0125* (ENSG00000226777), *WDR76* (ENSG00000092470), *AC005544.1* (ENSG00000214167), *KIAA1328* (ENSG00000150477), and *LL22NC03-63E9.3* (ENSG00000220891).

I also checked whether I find evidence for transcription of the AD-associated upregulated ncRNA *51A* that has been recently identified (see Section 6.1) [424]. I indeed observed a differentially expressed probe that is upregulated in AD ($p \sim 0.002$, $q \sim 0.11$) mapping antisense to the *SORL1* gene but it is not located antisense to intron 1 of the gene.

Recently, the lncRNA *LINC00299* has been found to be disrupted in subjects with neurodevelopmental disabilities, adding yet another example to the set of ncRNAs that may play a significant role in human developmental disorders [445]. Although *LINC00299* was not found to be classified as differentially expressed in the presented analysis, a total of six probes out of the 32 sense probes that were designed for *LINC00299* had a q-value of smaller than 0.2. However, due to the strict quality criteria, all these six probes were filtered and therefore not considered in subsequent analyses. It is therefore possible that *LINC00299* is a false negative, particularly because lncRNAs sometimes have very low expression values and may therefore be barely distinguishable from the background expression level.

I also characterized the functionality of differentially expressed loci more systematically by a GO terms enrichment analysis. The analysis was done separately for each of three available GO domains[5]:

- "Biological process": operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms
- "Molecular function": the elemental activities of a gene product at the molecular level such as binding or catalysis
- "Cellular component": the parts of a cell or its extracellular environment

To identify enriched GO terms, I used GOrilla (last updated on August 3, 2013) [446, 447] and GO-TermFinder [448] but I hereafter focus on the results of the latter. For summarization and visualization, I used REVIGO [449].

I performed a GO terms enrichment analysis separately for putative protein-coding and non-coding loci. For pseudogenes, no enrichment analysis was possible due to the small number of differentially expressed pseudogenes and their unknown function. Because I integrated multiple sources for non-coding loci, I could only use those that overlapped with annotated *Gencode* v14 genes (321 out of 460). For non-coding loci originating from introns (e.g., as identified by Nakaya et al. [429]),

[5] taken from ftp://ftp.geneontology.org/go/www/GO.doc.shtml, last accessed in August 2013

I used the ID of the surrounding gene if available regardless of its class (e.g., coding or non-coding). For the background population, I used all annotated *Gencode* v14 genes that overlapped in sense direction with any of the 113,047 probes that passed all quality filters.

In summary, the results revealed a diverse set of specific dysregulations. Most of the enriched GO terms had known AD-associations, and representative and dominant subsets include the following (Figure C.2 and Figure C.3 for a full list):

- **Biological process**: fibrinolysis, glycoprotein biosynthesis, cellular response to acid (including cell cycle), collagen fibril organization, regulation of blood coagulation, extracellular matrix organization, L-serine metabolism, anatomical structure formation involved in morphogenesis
- **Molecular function**: phospholipase inhibitor activity, calcium-dependent phospholipid binding, glycine hydroxymethyltransferase activity, lipoteichoic acid receptor activity, Rab GTPase binding, ryanodine-sensitive calcium-release channel activity, phospholipase inhibitor activity, beta-galactoside alpha-2,6-sialyltransferase activity, calcium ion binding, transcription regulatory region DNA binding
- **Cellular component**: extrinsic to (plasma) membrane, early endosome, myelin sheath adaxonal region, myelin sheath, (proteinaceous) extracellular matrix, extracellular region

6.2.7 Characterization of Differentially Expressed Loci Associated with Chromatin and Epigenetic Stability

I found a large number of differentially expressed loci with known or unknown chromatin-associated functions (Table 6.2 and Table 6.3, respectively) as well as functions for maintaining epigenetic stability. These are described in more detail below.

6.2.7.1 Chromatin-Associated Loci with Known Functions

The GO enrichment analysis identified an enrichment of the GO term 'histone deacetylase regulator activity" (GO:0035033). Such altered activity of histone deacetylases may be mediated, for example, by *TP53*, which is found to be upregulated in AD. It is known that *TP53* may positively regulate histone deacetylation, among the myriad of functions in response to cellular stress. Intriguingly, it has been argued that hypoacetylation may already be sufficient to trigger apoptosis [402].

Table 6.2: *Functionally characterized differentially expressed loci with known chromatin-associations. For column explanations, see Table 6.1. Note that because functions for non-coding genes are largely unknown, the list contains only loci originating from functionally characterized protein-coding genes.*

(Surrounding) Gene	*Ensembl* ID	Signal origin	Reg. in AD	n_{diff}
BDNF	ENSG00000176697	Exon of protein-coding gene	−	2
KIF20A	ENSG00000112984	Exon of protein-coding gene	+	2
MKI67	ENSG00000148773	Exon of protein-coding gene	+	2
PPARG	ENSG00000132170	Intron of protein-coding gene	+	2
TCF7L2	ENSG00000148737	Intron of protein-coding gene	+	2
ZBTB8B	ENSG00000215897	Exon of protein-coding gene	−	2
BARX2	ENSG00000043039	Exon of protein-coding gene	+	1
BRE	ENSG00000158019	Intron of protein-coding gene	−	1
C21orf7	ENSG00000156265	Exon of protein-coding gene	+	1
CASC5	ENSG00000137812	Exon of protein-coding gene	+	1
CEP55	ENSG00000138180	Exon of protein-coding gene	+	1
CHD8	ENSG00000100888	Intron of protein-coding gene	−	1
CUX1	ENSG00000257923	Intron of protein-coding gene	+	1
ESCO2	ENSG00000171320	Exon of protein-coding gene	+	1
HES2	ENSG00000069812	Exon of protein-coding gene	+	1
HIST1H3I	ENSG00000182572	Exon of protein-coding gene	+	1
HMGA2	ENSG00000149948	Intron of protein-coding gene	+	1
ISL1	ENSG00000016082	Exon of protein-coding gene	+	1
MSH3	ENSG00000113318	Intron of protein-coding gene	−	1
MTF2	ENSG00000143033	Exon of protein-coding gene	−	1
NSD1	ENSG00000165671	Intron of protein-coding gene	−	1
RUNX2	ENSG00000124813	Intron of protein-coding gene	+	1
SHMT2	ENSG00000182199	Exon of protein-coding gene	+	1
SMAD6	ENSG00000137834	Intron of protein-coding gene	+	1
SMARCA5	ENSG00000153147	Intron of protein-coding gene	−	1
TFAP2A	ENSG00000137203	Exon of protein-coding gene	+	1
TP53	ENSG00000141510	Intron of protein-coding gene	+	1
TP53	ENSG00000141510	Exon of protein-coding gene	+	1
YAP1	ENSG00000137693	Intron of protein-coding gene	+	1

Among the differentially expressed loci, I also identified a number of other genes with known chromatin-associations (Table 6.2). A few of them are particularly noteworthy and are described in more detail in the following[6]:

- **BDNF** has crucial functions in neuron survival, growth and differentiation of neurons and synapses, long-term memory, and chromatin regulation [434, 435]. I observed downregulation in AD, which is consistent with previous findings. The activity of *BDNF* has been shown to be under strong epigenetic control via DNA methylation, chromatin-modifying enzymes and the microRNA machinery (reviewed in [450]). Because splicing isoforms in particular are controlled through epigenetic mechanisms, the observed downregulation of *BDNF* may thus be a direct consequence.

- **HMGA2** is implicated in a myriad of processes such as apoptosis, mitosis, cell signaling, DNA damage, chromosome condensation, and regulation of transcription. *HMGA2* encodes a protein that can phosphorylate histone H2A at position 139 (H2AS139, see Section 2.1.4.1 for a functional description of this histone PTM), with additional functional roles in heterochromatin assembly and chromatin organization. Intriguingly, *HMGA2* has been found to contribute significantly to brain volume [451].

- **ESCO2** encodes a protein with acetyltransferase activity and has recently shown to be required for cohesin acetylation in pericentric heterochromatin and more generally chromosome segregation [452]. Expression of *ESCO2* is cell cycle-dependent and only transiently expressed during S-phase (reviewed in [452]).

- **C21orf7** encodes a protein that is implicated in the Mitogen-activated protein kinases signaling cascade. *C21orf7* interacts with *GPS2* and is an integral subunit of the *NCOR1-HDAC3* complex, which is a critical epigenetic regulator for circadian clock genes [453].

- **CHD8** encodes a DNA helicase with chromatin remodeling functions. It acts as a transcription repressor (negative regulator of the Wnt signaling pathway, suppressor of *STAT3* activity and *TP53*-mediated apoptosis) and also interacts with *CTCF*. Recently, it has been shown that in order to carry out its inhibitory effect, *CHD8*-mediated recruitment of histone H1 to Wnt target genes is essential [454].

- **SMARCA5** has helicase and ATPase activities and may therefore alter chromatin structure by nucleosome remodeling (see Section 2.1.4.4). It is also implicated in the cell cycle; specifically, it is involved in the replication of pericentric heterochromatin and the maintenance of chromatin structures during DNA replication. Furthermore, it is part of various chromatin remodeling complexes such as the *WICH* complex. It is also associated with various histone

[6] The following gene descriptions have in part been taken from *Entrez Gene* (`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene`, last accessed in August 2013) and *UniProtKB/Swiss-Pro* (`http://www.uniprot.org/`, last accessed in August 2013)

PTMs such as H2AXS142ph (which it mediates) and histone deacetylation (by interacting with histone deacetylase 2 [455]).

- **NSD1** encodes a histone methyltransferase that preferentially methylates H3K36 and H4K20 (see Section 2.1.4.1 for a functional description of these histone PTMs). It is also associated with metal ion binding (zinc). Dysregulation of *NSD1* has furthermore been repeatedly implicated in various diseases and cancer types [132] but direct links between *NSD1* and AD have yet to be established.
- **MTF2** encodes a *Polycomb* group protein that specifically binds H3K36me3 and subsequently recruits the *PRC2* complex, resulting in increased H3K27me3. It also regulates the transcriptional program during cell differentiation. It, however, may also act as local inhibitor of *PRC2* activity. Similar to *NSD1*, it also has zinc ion binding activity.

In summary, I identified various genes with histone-modifying functions (H2AS139ph — *HMGA2*; H2AXS142ph — *SMARCA5*; H3K27me3 — *MTF2*; H3K36me — *NSD1*; H4K20me — *NSD1*; histone acetylation — *ESCO2*; histone deacetylation — *C21orf7*) and histone PTM readers (H3K36me3 — *MTF2*; methylated histone residue binding — *CHD8*; histone acetylated lysine binding — *SMARCA5*).

Using `GOrilla` based on a gene list that has not been further filtered (see last paragraph in Appendix C.2), I also found a weak significant enrichment of proteins involved in chromatin assembly (*HMGA2*, *TP53*, *SMARCA5*) and chromatin binding (*PPARG*, *NSD1*, *TCF7L1*, *TCF7L2*, *WHSC1*, *SUV39H1*, *HMGA2*, *SHMT2*, *CUX1*, *SMAD6*, *YAP1*, *TP53*, *SMARCA5*, *RUNX2*). Although the term "chromatin assembly" did not show up as significantly enriched with the more stringent filtered list, these results suggest that altered regulation of genes implicated in chromatin assembly and binding may also play important roles in the etiology of AD.

6.2.7.2 Chromatin-Associated Loci with Unknown Functions

I also found a number of loci with known chromatin-associations but uncharacterized functions. In Table 6.3, I summarize differentially expressed loci from Mondal et al. [188] and Khalil et al. [201] that did not overlap with known genes (according to *Gencode* v14) and therefore have unknown functions. However, their chromatin-association has been experimentally verified.

Table 6.3: *Functionally uncharacterized differentially expressed loci with known chromatin-associations. Only loci that did not overlap with known genes (according to* Gencode *v14) and that have been identified by Mondal et al. [188] and Khalil et al. [201] are shown. For column explanations, see Table 6.1.*

Location	Reg. in AD	Source	n_{diff}
chr10:+127758121–127963700	+	Mondal et al. [188]	9
chr3:+123380693–123498543	+	Mondal et al. [188]	4
chr9:+21506439–21507378	+	Khalil et al. [201]	2
chr14:+101447501–101448399	−	Mondal et al. [188]	2
chr18:+459593–477126	+	Mondal et al. [188]	2
chr21:+36173450–36255426	+	Mondal et al. [188]	2
chr1:+98459693–98460854	−	Khalil et al. [201]	1
chr2:+238235835–238304366	+	Mondal et al. [188]	1
chr5:+39416553–39421200	+	Mondal et al. [188]	1
chr5:+172745556–172752745	+	Mondal et al. [188]	1
chr6:+43828988–43829296	+	Khalil et al. [201]	1
chr6:+98429073–98429545	+	Khalil et al. [201]	1
chr6:+169621368–169653239	+	Mondal et al. [188]	1
chr8:+130696593–130700510	+	Khalil et al. [201]	1
chr11:+72447205–72450376	+	Khalil et al. [201]	1
chr11:+86625764–86626534	+	Khalil et al. [201]	1
chr11:+122011788–122012636	+	Khalil et al. [201]	1
chr13:+111545560–111557540	−	Mondal et al. [188]	1
chr17:+57719–58170	+	Khalil et al. [201]	1
chrX:+68299562–68301518	+	Khalil et al. [201]	1

6.2.7.3 Loci Implicated in the Cell Cycle and Epigenetic Stability

Typically, vertebrate neurons are amitotic (with few exceptions such as sensory neurons of the olfactory epithelium) and the cell replication machinery therefore seems to be switched off. Intriguingly, however, in AD, it is now well established that vulnerable neurons display aberrant re-entry into the cell cycle despite their terminally differentiated status [456–458]. Altered cell cycle regulation seems to be indeed particularly pronounced in AD because I similarly found at least 27 differentially expressed genes with known functional roles in the cell cycle (*TCF7L2*, *TP53*, *HMGA2*, *SMAD*6, *KIF20A*, *MKI67*, *CEP55*, *CASC5*, *ESCO2*, *NCAPG*, *ASPM*, *WDR11*, *TGFBR1*, *CCNB2*, *KIF23*, *RRM2*, *APOBEC3G*, *DBC1*, *PRKCE*, *SEPT11*, *NUF2*, *PARD3B*, *SEPT9*, *MCM8*, *HNF4A*, *NEDD9*, *RALA*). Additionally, I identified various differentially expressed probes located in antisense direction to genes with cell cycle (regulatory) functions. For example, I identified an antisense probe to the gene *RCC1* (ENSG00000180198), which plays a key role in mitosis and is implicated in the regulation of chromosome condensation in the S phase of the cell cycle. It also binds to both nucleosomes and double-stranded DNA.

Importantly, as shown in Chapter 2 and Chapter 4, DNA replication and mitosis more generally have crucial roles for epigenetic stability. For example, gradual accumulation of errors during the lifetime of a cell caused by the inability to precisely restore the parental modification pattern (Figure 4.7) may quickly exceed the tolerable error threshold, therefore ultimately resulting in non-negligible changes in the transcriptional program. Typically, neurons are long-lived and therefore permanently maintain a precisely defined transcriptional state [64]. Vulnerable neurons in AD, however, re-enter the cell cycle and therefore it seems plausible to hypothesize that this may cause epigenetic instability that initiates a cascade of downstream effects. This seems to be particularly the case if the abundance and activity of chromatin-modifying enzymes that are crucial for restoring the premitotic histone modification pattern are also altered (e.g., HATs and HDACs).

In agreement with this, I also observed upregulation of the gene *HIST1H3I* (ENSG00000182572), which encodes a member of the histone H3 family. Histones are among the most abundant proteins in the cell, and their cell cycle-dependent gene expression is well-established. It therefore appears plausible that dividing neurons have an increased need for newly synthesized histones. Indeed, H3 seems to play a central role for learning and memory function (reviewed in [400]). In AD, altered histone H3 homeostasis may be caused by soluble Aβ, which can act as a powerful signaling molecule [400]. Additionally, cytoplasmic histone H3 has been shown to be subject to increased phosphorylation, which may contribute to neuronal dysfunction and neurodegeneration in AD [459].

6.2.8 Alzheimer as an Evolutionarily Young Disease

To address the hypothesis that AD is an evolutionarily young disease, I analyzed the conservation of AD-associated genes (both protein-coding and non-coding) by employing splice maps based on multiple alignments [115]. I chose this method because the evolutionary histories of ncRNAs have been notoriously hard to study due to their low level of sequence conservation that precludes comprehensive homology-based surveys and makes them nearly impossible to align. However, the conservation of gene structure and particularly the conservation of splice sites may also be used to establish homology [115]. Splice sites therefore leave phylogenetic footprints, and conserved patterns of splice sites may be used to predict novel transcripts from multiple genome alignments (hereafter called splice maps). This has been performed successfully and repeatedly in the past (reviewed in [115]).

I now briefly explain the methods, and I refer to Appendix C.4 for more details. The splice site analysis was done in collaboration with Anne Nitsche[7]. In summary, we compared the conservation of the differentially expressed protein-coding and non-coding genes that are annotated in *Gencode* v14 (147 genes with multi-exonic transcripts and a total of 3,354 splice sites and 122 genes with

[7]Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, anne@bioinf.uni-leipzig.de

multi-exonic transcripts and a total of 1,249 splice sites, respectively) for 35 mammalian species (human to opossum).

For both differentially expressed protein-coding and non-coding genes, we first generated a splice map that contained the exact aligned genomic coordinates for all available species and the corresponding `MaxEntScan` score for each listed splice site. For comparison and for assessing the significance of the observed values, we also constructed corresponding background datasets based on all Gencode v14 annotated protein-coding and non-coding genes, respectively. Based on human as reference, for each annotated gene, we then compiled a list of all splice sites located at their exon/intron boundaries. For each of the other mammalian species, we then calculated the fraction of "conserved" splice sites (according to various criteria, see Appendix C.4). This procedure therefore allowed us to determine gene conservation between pairs of species. One advantage of this approach, as compared to alternative approaches that do not work on the gene level but rather on the splice site level directly, is that it weights each gene equally regardless of its length and the number of splice sites it contains. We considered a gene between two species as conserved if a particular percentage of the splice sites located in that gene (hereafter denoted gene conservation threshold c) was conserved. Because c is expressed as a percentage, it therefore makes comparisons among genes of unequal length possible. To examine the effect of the exact value of c, we varied c ($c > 0\%$, $c > 10\%$, $\ldots$,$c > 90\%$, $c = 100\%$). Using the corresponding background splice maps, it was then possible to measure the evolutionary conservation of differentially expressed protein-coding and non-coding genes and to assess their statistical significance.

For protein-coding genes, we expectedly found that conservation is generally very high among mammals. Independent of the gene conservation threshold, we found only negligible conservation differences between AD-associated and all protein-coding genes and therefore no indication for stabilizing selection (although some differences were statistically significant).
For non-coding genes, we found that conservation in general decreased rapidly with the evolutionary distance (Figure 6.6). This was particularly apparent for non-primate mammals and $c > 40\%$. However, we found evidence for stabilizing selection for AD-associated non-coding genes (AD-ncRNA genes) because independent of the specific gene conservation threshold, AD-ncRNA genes had higher conservation values as compared to all ncRNA genes, with significant differences particularly in closely related species (primates). The higher the gene conservation threshold, the less pronounced the differences were in more distantly related species, whereas the differences in closely related species remained.
Collectively, the findings are in agreement with the hypothesis that AD-ncRNA genes are evolutionarily young because they were subject to extensive recent changes in their gene structure. However, extremely low conservation values (particularly if they are very different from the values from closely related species) have to be treated with caution (see Discussion).

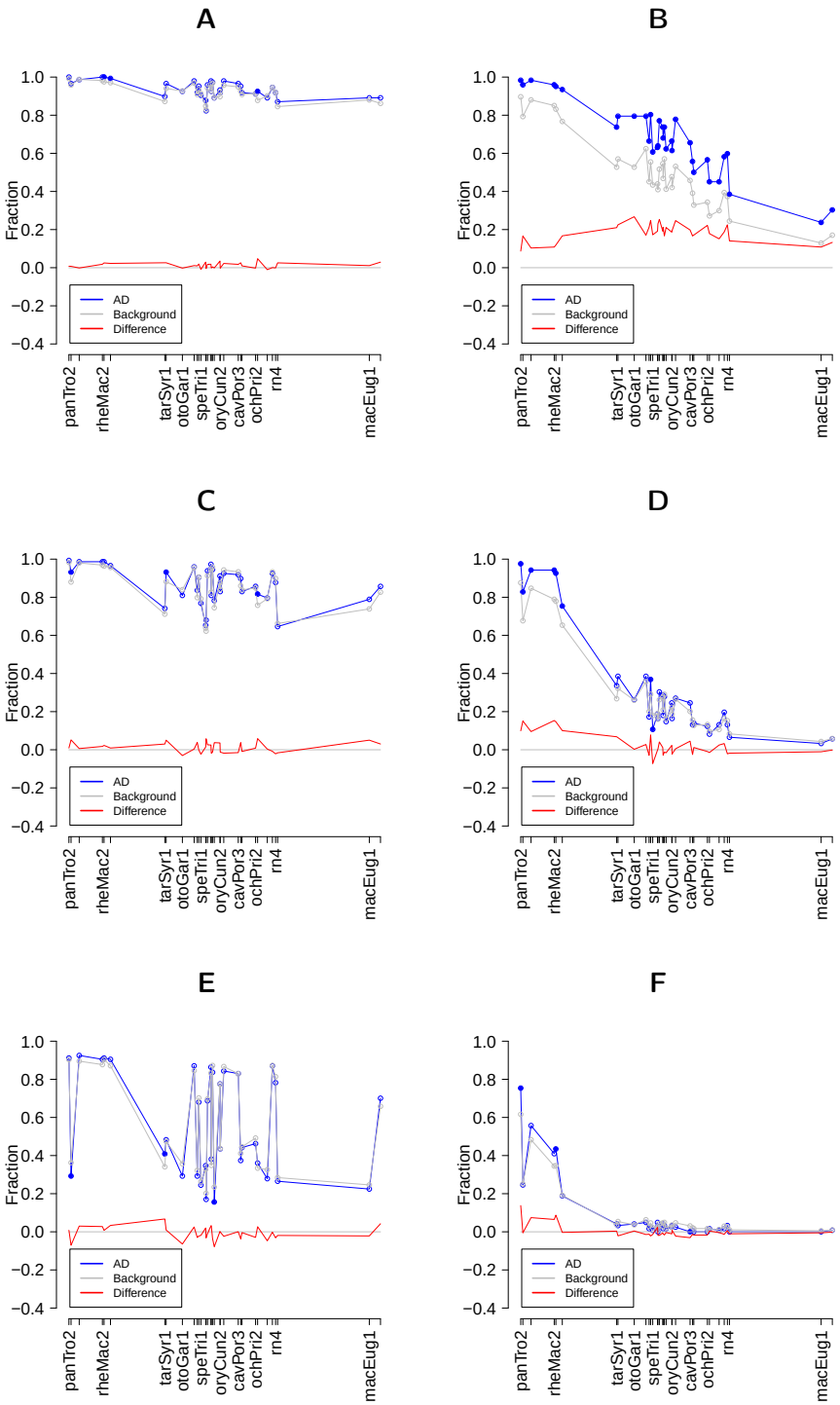


Figure 6.6: *Results of the splice site conservation analyses based on differentially expressed genes with different values for the gene conservation threshold c. The fraction of conserved genes across various mammals from chimpanzee (`panTro2`) to opossum (`monDom5`) is shown as compared to human. Left (A,C,E): protein-coding genes. Right (B,D,F): non-coding genes. (A,B): $c > 0\%$; (C,D): $c > 40\%$; (E,F): $c > 80\%$. Filled dots indicate significant differences.*

6.3 Discussion

In this chapter, I characterized AD-associated genomic loci using the `Alzheimer Custom Array`. To identify differentially expressed probes and genes and for their quantification and initial characterization, I used rigorous bioinformatics approaches. A distinctive feature of the presented study is that a large variety of caRNAs and ncRNAs more generally are deliberately included on the `Alzheimer Custom Array`. Although their disease-relevance is increasingly being recognized (see Section 2.1.4.3), previous systematic gene expression profiling studies nevertheless focused predominantly on protein-coding genes (reviewed in [410]). Consequently, so far, only individual AD-associated ncRNAs have been identified and functionally characterized (e.g., [424, 441]).

I identified a total of 764 differentially expressed genomic loci that belong to very different classes. Due to the composition of the `Alzheimer Custom Array`, the majority of them are putative ncRNAs. A preliminary functional analysis using the `GO` database revealed that dysregulations in AD are numerous and diverse, which is in agreement with previous work (reviewed in [410]). However, due to the complexity of any gene expression profiling study, the results presented here provide only a first glimpse into the data. For example, although performing a `GO` enrichment analysis is useful to identify general patterns in the data, to generate hypotheses, and to some extent also to assess the validity of the underlying methodology, further analyses and studies based on the enrichment results are certainly required.
Follow-up analyses include, but are not limited to: further characterizing putative novel antisense transcripts and ncRNA in general, attempts to identify particular splicing variants that are differentially expressed rather than full genes, and importantly also experimental validation for selected candidates. The latter is already ongoing and results are expected soon. Also, differentially expressed loci with unknown transcript structure can be further analyzed by integrating information from additional sources (e.g., `tblastx` or `RNACode` [460] to test for coding potential). Further characterization of identified differentially expressed pseudogenes also seems worthwhile, particularly because little is known about their AD significance and precise biological function. Indeed, increasing evidence suggests that pseudogenes may have special regulatory functions [461–464].

To identify differentially expressed loci, I employed a novel two-step procedure that aims to reduce false positive and false negative rates. Only loci that passed both the q-value filter (step 1) and the binomial test (step 2) were further considered. The results of the `GO` terms enrichment analyses suggest that this approach worked well because the vast majority of genes with functional characterizations have immediately apparent AD associations. Clearly, however, this is only an indication, and it should be further strengthened by experimental validation (see above).

I explicitly addressed two hypotheses in this chapter. First, I asked whether the genomic loci that are differentially regulated in AD are evolutionarily young (i.e., whether they show signs of recent changes in their genomic structure). I expected that this should be particularly evident for

non-coding regions and only to a much lesser extent for protein-coding regions because the latter have a high evolutionary pressure and are therefore largely conserved across mammals. Estimating the evolutionary age of lncRNAs is generally non-trivial because they have been notoriously hard to study due to their low level of sequence conservation that precludes comprehensive homology-based surveys and makes them nearly impossible to align. However, Nitsche et al. [115] showed that lncRNAs have a fast turnover of their intron/exon structures although they are evolutionary ancient components of vertebrate genomes. Estimating their evolutionary age by analyzing the conservation of splice sites is therefore an appropriate method. However, alignment quality and completeness may have a substantial influence on the conservation results. Alignment quality issues generally may arise from alignment gaps as a consequence of incomplete or erroneous genome annotations or inaccuracies of the multiple alignment in regions of splice junctions, for example, which was particularly an issue for non-model organisms (see Appendix C.4). As a control, I therefore also calculated the percentage of positions that can be aligned (independent of any splice site conservation, see Figure C.4) and the fraction of conserved genes among alignable genes (Figure C.5). This revealed that conspicuously low gene conservation values for particular species with respect to their closely related species are often caused by a low fraction of alignable genes. Thus, extreme values in Figure 6.6 have to be treated with caution and should be interpreted only in conjunction with the control measures. Nevertheless, the results indicate that differentially expressed non-coding loci are indeed evolutionarily young and particularly exposed to changes in their exon/intron structure in the primate lineage. For protein-coding loci, however, I expectedly observed only negligible differences.

The second hypothesis I addressed concerned the significance of chromatin in AD. Specifically, I explored the possibility that a dysregulated CC may play important roles in AD. Generally, dysregulation of the CC may be achieved by differential expression of chromatin-associated loci or by altered epigenetic control of AD-associated genes. Chromatin-associated loci come in many facets. They may, for example, encode the most fundamental components of chromatin such as histone proteins or chromatin-modifying enzymes or integral parts of the chromatin-modifying enzyme complexes. In addition, they may encode molecules that guide chromatin-modifying complexes to particular genomic loci (e.g., lncRNAs) or proteins that directly or indirectly influence epigenetic stability (e.g., by altering cell cycle regulation and therefore the stability of epigenetic states). In support of the hypothesis, I identified chromatin-associated, differentially expressed loci for all of these four general classes. For example, I found that various genes encode proteins that directly or indirectly, via association with additional proteins, constitute histone reader (e.g., H3K36me), writer (e.g., H2AS139ph, H2AXS142ph, H3K27me3, H3K36me, H4K20me, unspecific histone acetylation) and eraser functions (e.g., unspecific histone deacetylation).

In agreement with previous work, I found aberrant evidence for altered cell cycle regulation in AD. Although AD-associated re-entry into the cell cycle seems to precede amyloid-ß plaques and neurofibrillary tangles formation and therefore is more likely to be a consequence rather than a

cause of the disease [457, 459, 465], it has important implications for epigenetic stability. Neurons are typically long-lived and maintain a precisely defined transcriptional state for a long time [64] but gradual accumulation of errors caused by DNA replications may constitute the final straw towards an irreversible and uncontrollable cascade of downstream effects that ultimately leads to apoptosis. Such transcriptional dysregulation is furthermore exacerbated by unbalanced activities of chromatin-modifying enzymes such as HATs and HDACs [466].

Consistent with the finding of a large number of differentially expressed loci with chromatin associations, both chromatin-associated datasets and almost all cell cycle tiling array regions that were included on the `Alzheimer Custom Array` were also enriched on the probe level. The feature enrichment analysis also revealed a general tendency for enrichment of probes located in antisense direction to known transcripts and enrichment of a diverse set of ncRNAs such as snoRNAs, lncRNAs, and pseudogenes. Particularly the finding of enrichment of lncRNAs is in agreement with the dysregulated CC hypothesis, due to the guiding function of lncRNAs for chromatin-modifying complexes. Indeed, it is increasingly recognized that particularly lncRNAs often associate with chromatin regulators [467] (see also Section 2.1.4.3). However, in contrast to protein-coding genes, functional annotation of ncRNAs is only in its infancy and so far, only several dozen candidate ncRNAs have been functionally characterized. Similarly, automated annotation attempts are only beginning to emerge (e.g., `Linc2GO` [468]). Therefore, it is not particularly surprising that the `GO` terms enrichment analysis was largely devoid of immediately apparent chromatin associations because a non-negligible part of the differentially expressed loci may correspond to functionally uncharacterized chromatin-associated transcripts.

The feature enrichment analysis revealed a general depletion of probes located in introns in sense direction, whereas probes located antisense to introns, particularly introns that were previously identified to harbor ncRNAs [429], were enriched. Indeed, transcripts located in antisense direction to introns seem to be of increasing significance. For example, in AD, two disease-associated ncRNAs have recently been identified [52, 424], and it seems plausible that this is a much more pervasive phenomenon than currently recognized. Thus, the data suggest that intronic transcription in general is not globally enriched but it instead occurs predominantly in particular introns. This would explain both the depletion and enrichment result as described above.

Aberrant transcription within introns may be a consequence of a differential chromatin structure. This is consistent with the findings of Zhu et al. [469], who measured chromatin states for a total of seven common histone PTMs for a large and diverse set of human tissues, including cells from different brain sections. The authors found that cells from brain sections display particularly strong evidence for preferential utilization of intronic regulatory elements. Thus, in brain cells, a highly specialized and restrictive chromatin structure may facilitate access to and consequently transcription from intronic regions. In agreement with this, genes with "a high density of conserved noncoding sequences in their introns are expressed at higher levels in brain and enriched for functional

annotations related to neuronal physiology" [469, p. 648]. Thus, many of the identified differentially expressed non-coding loci may have important chromatin-associated functions that are implicated in the etiology of AD, either directly or indirectly. In summary, I find various indications that a dysregulated CC plays important roles in AD although it is very difficult to establish causalities due to the complexity of the disease (see below) and its largely unknown nature. Intriguingly, both the chromatin regulatory system and AD are evolutionarily young (see Section 3.1.3 for the former), and it is tempting to speculate that this is indeed not a coincidence.

Although recent studies indicate that microarray results are acceptably consistent among platforms and analysis techniques [410, 470], a number of other technical considerations and limitations are noteworthy for the interpretation of the results and as potential explanation for the failure of detection of differential expression of some known AD-associated genes. Generally, AD is extraordinarily difficult to study due to its complexity and heterogeneity. This is, for example, also exemplified by the functional diversity of GO enrichment terms. Indeed, as mentioned earlier, only a small percentage of AD cases can be linked to mutations in specific genes, and a large array of putative causes have been postulated. A frequent issue in AD studies is limited reproducibility, and except for a few AD hot spots that are frequently identified, different studies often have little overlap. However, gene expression profiling as well as bioinformatics analytical and evaluation tools have advanced considerably during the last few years [410]. In conjunction with increased scientific rigor, data is now more reproducible than before.

Limited reproducibility is at least partly caused by the following technicalities. Sample acquisition and preparation are crucially important, particularly for post-mortem tissue as used in this study. First, it is not clear whether the survived neurons obtained in final stages of AD are generally representative of AD in the disease. Second, the collected brain tissue may contain a mixture of different cell types that may reflect the AD pathophysiology to varying degrees [410, 471]. Third, samples from healthy individuals may be neuron-rich, whereas AD samples are typically neuron-depleted, therefore introducing potential biases in subsequent analyses [410]. More generally, the origin of the samples is also of relevance. For example, the temporal lobe-hippocampus and the prefrontal cortex seem to be particularly rich in differential gene expression [410, 472, 473]. The samples in this study are also from the temporal lobe.

Furthermore, RNA integrity is a major concern for any gene expression study because sufficiently high RNA quality is fundamentally important for reproducibility. RNA integrity numbers (RIN) are the current gold standard to assess RNA quality, and they have been determined for all of the samples. RIN values of post-mortem samples are typically very low due to natural RNA degradation. In this study, only samples with a RIN $\geq$ 5.0 entered the processing, with an average RIN value of 6.6. This is substantially larger than reported values for post-mortem brain samples. For example, Koppelkamm et al. [474] reported mean values of 2.8 for brain tissue.

Another complication in AD gene expression profiling studies is the degree of pathology of disease

samples. It is widely known that different stages in AD are neuropathologically recognizable and correspond to unique biochemical outcomes (Braak stages [475]). For example, synapse-related genes are upregulated in early stages and downregulated in later stages of AD [476]. Controlling for the disease stage is also important because it has been shown that the largest aberrant expression changes occur during progression from mild to moderate dementia [410, 477]. I predominantly included patients with severe AD (Braak stage 5 in particular) and therefore circumvented the need to explicitly control for disease stage.
Similarly, age is often a confounding variable because gene expression generally changes with age (e.g., age-related differences in the dysregulation of genes related to neuroinflammation [478]), and AD patients are frequently older than healthy individuals. In the identification of differentially expressed probes, this may introduce an additional bias unless explicitly controlled for, and I therefore included the variable age in the model.

Lastly, it is difficult to identify the most relevant differentially expressed probes or genes because the most significant ones (either by using p-values or fold changes) may not be most relevant from a biological and pathophysiological perspective [410]. This is particularly evident for regulatory structures such as TFs networks due to their ultrasensitivity (Figure 2.10) [479]. The correct differential detection of transcripts with low expression levels (e.g., lncRNAs) poses additional problems due to their proportionally increased expression variability and noise levels [410]. The two-step approach for the identification of differentially expressed genes may be particularly useful in that regard.

I addressed the hypothesis that AD is an evolutionarily young disease by analyzing splice site conservation. Intriguingly, the importance of (alternative) splicing for AD has steadily increased recently, and it seems to play a non-negligible role in the pathology of AD and neurodegeneration more generally [114, 441, 442, 480–482]. For example, Merkin et al. [483] established a link between alternative splicing and protein phosphorylatability. This therefore delimits the scope of kinase signaling pathways, which play important roles in AD and have also been identified in the `GO` terms enrichment analysis. As shown by Barbosa-Morais et al. [484], most splicing patterns in vertebrates are *cis*-directed. Variation of *in cis* splicing may be initiated, targeted and executed by ncRNAs, as exemplified by *51a* and *38A* [52, 424], both of which are upregulated in AD. Indeed, the vast majority of multi-exon genes undergo alternative splicing (reviewed in [109, 484]) but such transcript isoform variation is not captured by gene-level analyses [410]. Identification of splicing isoforms for complex transcriptomes is, however, challenging with microarray platforms, even with carefully designed CEMs (see Chapter 5). Expression measurements of individual exons may show large variability due to the decreased number of probes per exon as compared to the number of probes per gene for gene-level analyses, therefore making it difficult to reliably identify isoform differences. RNA-seq and next-generation sequencing more generally offer great potential in that regard, and the next years may advance the understanding of the significance and causality of alternative splicing for the pathogenesis of neurodegenerative diseases substantially.

Chapter 7

Conclusions and Outlook

Chromatin is a highly dynamic and incredibly complex structure with crucial roles for cell differentiation, transcription, and the heredity of gene expression patterns across cell divisions (i.e., epigenetic inheritance). The cellular memory capacity that it harbors is vital in the development of multicellular organisms and also underlies cell differentiation. A great variety of chromatin-modifying enzymes in general and histone-modifying enzymes in particular are able to read/recognize, write, and erase chromatin marks in a context-dependent and/or context-independent manner, thereby establishing a very dynamical and complex information processing system.

In this thesis, I elaborated on the relatively new notion that chromatin may be regarded as a biological computer and strengthened the idea that it provides a potent and universal "language" in which computer programs or biological procedures may be written. In Chapter 3, I analyzed selected biological building blocks of the chromatin computer and showed that histone PTMs constitute the main memory with a capacity of several hundred megabytes of writable information per cell. However, redundancy is fundamentally important for the chromatin computer due to its inherently stochastic nature and volatility, therefore lowering the real memory capacities. But what is the biologically required level of redundancy in the chromatin regulatory system? Is the existing level of redundancy comparable or substantially higher than the biologically required level due to the step-wise increase in complexity and computational power of chromatin during evolution? These exciting questions remain to be tackled explicitly.

Chromatin-modifying enzymes represent the execution unit of the chromatin computer and implement the logical and arithmetical operations in the form of rewriting rules. Using these rules, I demonstrated that the chromatin computer is, at least in theory, computationally universal and may therefore be used to calculate any computable function or algorithm more generally. The results are therefore in agreement with the notion that eukaryotic chromatin may be regarded as a molecular computer able to perform computations, both in a biological context but theoretically also in a strict informatics sense. Chromatin is therefore another representative of the growing number of non-standard computing examples.

However, the question remains whether the computational and memory capabilities of chromatin are only interesting from a theoretical point of view or if the cell indeed utilizes and requires them.

Clearly, composition and mode of action of the chromatin computer are a product of evolution. Inevitably, therefore, the rules that govern the biological computations (i.e., the reading, writing, and erasing of chromatin marks such as histone PTMs) also changed fundamentally during evolution to implement the increase in complexity and to remain adaptive to the environment. The results in Chapter 4 suggest, however, that evolving a system of enzymes that can maintain a particular chromatin state roughly stably may be a relatively easy task. Epigenetic inheritance seems to require computational capabilities but if they are utilized by the cell to their full extent remains an open and extraordinarily interesting and relevant question.
Another interesting question is whether the computational power of chromatin may be exploited for artificially designed systems. It has previously been suggested that a chromatin computer may be more widely applicable for a variety of problems than a DNA computer. In addition, a chromatin computer may provide solutions that are easier to implement and it therefore may be better suitable for general-purpose programs. However, its inherently stochastic nature, volatility, complexity and the interdependency among its individual components raise the question whether it will be possible to design algorithms that produce trustworthy results.

In Chapter 4, as an example for the computational power that is harnessed in real biological systems, I formulated epigenetic inheritance as a computational problem. Epigenetic inheritance (see Chapter 2) is generally characterized by a high underlying complexity and an intricate interplay of its individual components. To address the complexity systematically and to identify the major players of the system and their interdependencies, I developed a flexible and chemically accurate stochastic simulation system for the study of recomputation-based epigenetic inheritance of individual histone PTMs. Theoretical analyses using rigorous computational models as performed in this thesis are important contributions that have already provided plenty of new insights and perspectives and will continue to do so. Because faithful propagation and reconstruction of patterns of histone PTMs across cell divisions may be solved with sufficient stability and accuracy by the chromatin computer, propagation of patterns of histone PTMs can therefore indeed be interpreted as a computational problem that is achievable through a small collection of rewriting rules. These rewriting rules are abstractions of a well-described class of enzymes and enzyme complexes combining reader, writer, and eraser domains for specific histone PTMs.

The finding that patterns containing patches of unmodified nucleosomes are more difficult to inherit than modified ones due to the ambiguity of the unmodified state (i.e., unmodified from the beginning versus information loss) raises the question of biological relevance. As argued in Chapter 4, although the absence of a signal may also be informative, it is tempting to speculate that inheriting the unmodified state is indeed not particularly relevant. Nevertheless, future work is necessary to shed light on the significance of this result.

I emphasize again that the focus in this thesis is on the computational task of reconstructing complex patterns of histone PTMs after DNA replication that is typical for somatic cells. In particular, I

do not claim that epigenetic inheritance across the germ line follows the same paradigm because information inherited through the germline for an effectively infinite number of generations is subject to Eigen's error threshold. As shown in Chapters 2 and 4, epigenetic inheritance mechanisms are less faithful as compared to genetic inheritance mechanisms, and consequently, the amount of stably inheritable epigenetic information is severely limited. Consistent with this, most, if not all, of the extraneous epigenetic information is erased during spermatogenesis and oogenesis. In contrast, the error threshold does not preclude inheritance of complex patterns of histone PTMs in somatic cell lines because the number of generations is limited, and usually small. Here, the degradation of the epigenetic information is acceptable for a while and counteracted by multiple layers of redundancy. However, it may eventually also relatively quickly lead to daughter cells whose epigenetic patterns are damaged beyond repair. To some extent, this effect may thus constitute an epigenetic version of aging.

Whether histone PTMs and the presence of histone variants are a cause or consequence of the transcriptional status is still hotly debated. In this thesis, I am completely agnostic about this issue since it has no impact on the conclusions. The programs that run on the chromatin computer (i.e., the schedules and concentrations of rewriting enzymes) are externally specified in the model. In particular, I make no statement in regards to whether the gene expression program is a direct consequence of, or at least dominated by, the chromatin state or whether it is entirely determined by classical transcription factor networks that are largely or even completely independent of the chromatin state. In computer science terms, I employ a model of computation that strictly distinguishes between (gene expression) programs and (histone modification) data.
It appears natural, in a next step, to remove this distinction and to ask whether chromatin itself can "learn" to reprogram itself, by making the gene expression programs an intrinsic function of the histone PTM data. Although this may be an extreme model that implicitly views transcription factors as being enslaved by histone PTM states at their gene loci, it is an important limiting case given that gene expression is clearly not independent of chromatin state.

Indeed, transcription and the dynamics of histone PTMs are tightly linked. Although a number of distinct mechanisms evolved that anchor chromatin modifications to the underlying DNA sequence, the chromatin regulatory system is neither completely determined by the underlying DNA nor fully detached from it. It has long been speculated that this semi-independence may be a common source of pathology that significantly or even etiologically contributes to diseases such as cancer and AD. As an example, I stressed the importance of the chromatin regulatory system in AD, the most common and irreversible form of dementia. The identification of numerous differentially expressed loci that belong to different classes of chromatin-associated transcripts indicates that dysregulation occurs in a heterogeneous, almost global fashion. Furthermore, the preferential utilization of intronic regulatory elements, which seems to be common principle in brain cells and AD in particular, have been suggested to be a consequence of a differential chromatin structure caused by chromatin-associated transcripts. I thus found good support for the hypothesis that a

dysregulated chromatin computer plays important roles in the etiology of AD although it is very difficult to establish causalities due to the complexity of the disease.
The finding of aberrant reactivation of cell cycle related genes is in agreement with the mounting evidence that suggests that in AD, the DNA replication machinery is indeed switched back on, which may ultimately accelerate the epigenetic aging of neurons. The progressive memory impairment in AD may therefore be a direct or indirect result of epigenetic processes. As shown in Chapter 4, several factors contribute to the task of recomputing the parental histone PTM state and thereby also influence the accumulation of errors in the wake of cell divisions. Altered regulation of genomic regions implicated in epigenetic stability may therefore easily maneuver the cell into an erroneous state that is difficult to recover from.

A multitude of open and exciting AD-related questions should be addressed more explicitly and systematically in the near future. First, for AD and age-related neurodegenerative disorders more generally, the prion hypothesis has recently gained experimental momentum. However, the connection to the overall and diverse dysregulation that occurs in AD is presently unclear. It remains to be seen whether the chromatin computer has a non-negligible role in this process. Second, it would be very interesting to determine which genomic regions and chromatin-associated transcripts in particular first show signs of dysregulation. This is particularly interesting because it is speculated that the pathological cascade of AD begins at least ten years before the appearance of the first clinical symptoms. Lastly, given the computational capabilities of chromatin, may it be possible to specifically reprogram the chromatin computer to counteract the detrimental and progressing effects of AD? Histone deacetylase inhibitors are promising candidates in that regard although it seems unlikely that they alone can reverse the globally altered expression landscape of AD.

For the identification of differentially expressed loci in AD (see Chapter 6), I designed a custom expression microarray. As shown in Chapter 5, high-quality microarrays for complex genomes require an appropriate and well-considered probe design strategy. The developed bioinformatics pipeline and the web server considerably automate and facilitate the design of custom expression microarrays, emphasizing in particular on target and probe selection and providing high flexibility for the selection and preprocessing of target sequences. Although microarrays in general are gradually replaced by superior technologies such as RNA-seq, they are still used ubiquitously for transcriptome profiling. Indeed, it has been shown repeatedly that microarrays and RNA-seq produce very similar and reproducible results. In addition, the technologies often even complement each other in transcriptome profiling. It can therefore be expected that microarrays will continue to be used for at least a few more years. Indeed, they can still provide novel biological insights, as evidenced by the `Alzheimer Custom Array`.

Lastly, the following question must be raised: What did we ultimately gain from considering chromatin as a molecular computer? Although this view may initially seem somewhat abstract and

artificial, it helped to make us realize at least the following four issues. First, it highlighted the complexity, ubiquity, and versatility of the chromatin regulatory system. In order to construct a simulation system, it also contributed to think about the rules of the system that ultimately execute a particular program, how they may be constructed, how complex they are, and how they are coordinated. Second, by the explicit comparison to ordinary, silicon-based computers, it emphasized that computation is not only an artificial idea but also a natural one and that chromatin is, at least in theory, computationally universal. In addition, it demonstrated that the chromatin regulatory system has striking analogies to amorphous computing, which has, to the best of my knowledge, not yet been observed before. Third, by analyzing its components and their interplay and by backtracking the evolution of the chromatin regulatory system, it stressed the fundamental importance of redundancy and memory in the system. Memory in particular allows the cell to keep a record of former states and to execute a specific program at any time point given a particular input (trigger). Thus, chromatin-associated changes may not be immediately phenotypically visible. In contast to TF networks, for example, the chromatin regulatory system is also not dependent on direct feedback mechanisms that actively maintain a particular state. Lastly and more generally, it helped to establish a deeper understanding of the capacities and limits of chromatin in general and somatic epigenetic inheritance in particular.

Appendices

Appendix A

Additional Details and Definitions for the Chromatin Computer

A.1 Details for the Memory Calculations for the Chromatin Computer and the Full Genome

In Table 3.2, I summarize the estimated information content (IC) of individual histones, nucleosomes and the total writable memory size of the full genome. Generally, the amount of information that can be stored in an entity represents its IC and is expressed in a unit of information, typically bits or multiples of bits (e.g., bytes, kilobytes, ...). In traditional information theory, one bit is typically defined as the uncertainty of a binary random variable x_b, under the assumption that the possible values 0 and 1 have equal probability [485]. Alternatively, one bit represents the information that is gained when the value of x_b becomes known.

For individual histone residues, the IC can be calculated as the logarithm to base 2 of the number of distinct states. Both the unmodified state and all types of modified states (e.g., methylation, acetylation, see Section 2.1.4.1 for a list) count as separate states because they may represent particular signals that are distinguishable by the cellular machinery. The IC of nucleosomes is then simply calculated as the sum of the IC of all residues, assuming that all residues may be modified independently from one another [12].

For the IC of core histones on the nucleosome scale, I summed over their individual IC, counting each of the four core histones twice. This was necessary because histones of the same type may indeed be modified distinctively, which can have functional consequences (see Section 2.1.4.4). For calculating the total memory capacity of the core histones and histone H1, I multiplied the IC of an individual nucleosome with the approximate number of total nucleosomes. For the latter, I used a value of 10 million, which accounts for the presence of nucleosome-free regions and was also used in previous calculations (e.g., [13]).

I calculated the values of the IC of individual histones as follows. For the lower limit, I used only histone PTMs that have been reported in humans. For this, I performed a comprehensive literature

search of reported human histone PTMs based on Tan et al. [116] (130, including PTMs to the linker histone H1.2), Xie et al. [117], the HIstome database [486] (only PTMs of the canonical histones and H1.2), Migliori et al. [487], Chen et al. [118] (only *in vivo* verified histone PTMs), Migliori et al. [487] (H3R2me2a and H3R2me2s), and Jack et al. [488] (H3K56me3). In total, I collected 223 distinct histone PTMs (191 for the four canonical histones H2A, H2B, H3, and H4 as well as 32 H1.2 PTMs). I, however, counted the different forms of histone methylation separately because they may indeed have distinct functions. Although the linker histone H1 is located outside of the nucleosome, it may also be post-translationally modified (see Section 2.1.2). The significance of histone H1 PTMs is still largely unexplored but evidence suggest that they also have important functions (see Section 2.1.2). To analyze the potential contribution of H1 for the memory of the CC, I therefore additionally included histone PTM on H1 in the memory size calculations.

For the calculation of the upper limit for the memory size, I first selected a reference sequence for each histone (Table A.1). Then, I determined the number of possible states (with respect to known types of histone PTMs) and correspondingly its IC at each amino acid residue for each of the four core histone proteins (Table 3.1).

I estimated the IC and total memory capacity for DNA based on two bits per base (corresponding to the four possible combinations of bases). I used a value of 200 bp for the length of one individual nucleosome, and the total memory size was calculated with 3,095,677,412 bp (i.e., ≈3 billion bp).

For DNA methylation, I used 1 bit per methylcytosine for the lower limit (absent versus present) and 2.3 bits per methylcytosine for the upper limit (absent versus one of the four cytosine derivatives that may be present and distinguishable by the cell: 5-methylcytosine, 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine; see Section 2.1.4.6). For the IC of DNA methylation for individual nucleosomes, I used a lower limit of 0 (no CpG dinucleotides and therefore no possible DNA methylation) and an upper limit of 200 (CpG island and therefore every base may be methylated if considered double-stranded). For the total memory capacity of DNA methylation, I was inspired by Lister et al. [241] who experimentally identified 62 and 45 million methylcytosines in *H1* and *IMR90* cells, respectively. As an approximation, I used a value of 50 million for the lower limit (which therefore approximately equals the average experimentally identified number of methylcytosines in the two cell types) and 100 million for the upper limit (to account for non-detected methylcytosines and more generally for the large temporal and spatial variation that has been observed for the frequency of DNA methylation).

Histone	Length (in amino acids)	IC (in bits)	*GenBank* accession number	Amino acid composition
H1	215	331.5	NP_005316.1	G:13, P:19, A:42, V:14, L:10, I:3, M:1, C:0, F:1, Y:1, W:0, H:0, K:57, R:4, Q:2, N:4, E:6, D:1, S:23, T:14
H2A	142	123.0	NC_000011.9	G:17, P:7, A:20, V:9, L:16, I:6, M:0, C:0, F:1, Y:4, W:0, H:3, K:13, R:12, Q:6, N:5, E:7, D:2, S:7, T:7
H2B	125	157.3	NC_000001.10	G:7, P:6, A:13, V:8, L:6, I:7, M:2, C:0, F:2, Y:5, W:0, H:3, K:20, R:8, Q:3, N:3, E:7, D:3, S:14, T:8
H3	135	137.6	M26150.1	G:7, P:6, A:18, V:6, L:12, I:7, M:2, C:1, F:4, Y:3, W:0, H:2, K:13, R:18, Q:8, N:1, E:7, D:4, S:6, T:10
H4	102	100.6	NC_000012.11	G:17, P:1, A:7, V:9, L:8, I:6, M:1, C:0, F:2, Y:4, W:0, H:2, K:11, R:14, Q:2, N:2, E:4, D:3, S:2, T:7

Table A.1: *Upper limit for the IC of histone H1 and the core histones that make up a nucleosome. For each histone class, a representative human amino acid translation was downloaded from* GenBank *(see accession number). The IC of a particular histone is calculated as the sum of the IC of its individual amino acids (Table 3.1). The last column summarizes the amino acid composition of the corresponding histones (i.e., the frequencies of each amino acid), with amino acids abbreviated as single letter codes.*

A.2 Formal Definition and Mode of Action of a Turing Machine

A Turing machine (TM) may be defined as a 7-tuple $< Q, \Gamma, \#, \Sigma, q_0, F, \delta >$:

1. Q: Set of states
2. Γ: Set of tape symbols
3. $\#$: A special blank symbol, $\# \in \Gamma$ and $\# \notin \Sigma$
4. Σ: Set of input symbols, $\Sigma \subset \Gamma$
5. q_0: Start state, $q_0 \in Q$
6. F: Set of final or accepting states, $F \in Q$
7. δ: Transition function $Q \setminus F \times \Gamma \rightarrow Q \times \Gamma \times \{\mathtt{L}, \mathtt{S}, \mathtt{R}\}$

Q, Γ and Σ must all be finite and non-empty. The same typically applies to F although it may be empty (TM that accepts no strings). Upon execution of a rule, we also allow the head to *stay* (S) at its current position. This is in contrast to standard definitions of a TM, for which the head moves either to the left or to the right but never stays at the same location. However, TMs with the *stay* option are equivalent to standard TMs and therefore computationally not more powerful. I adjusted the definition for practical reasons (e.g., histone-modifying enzymes may also stay bound

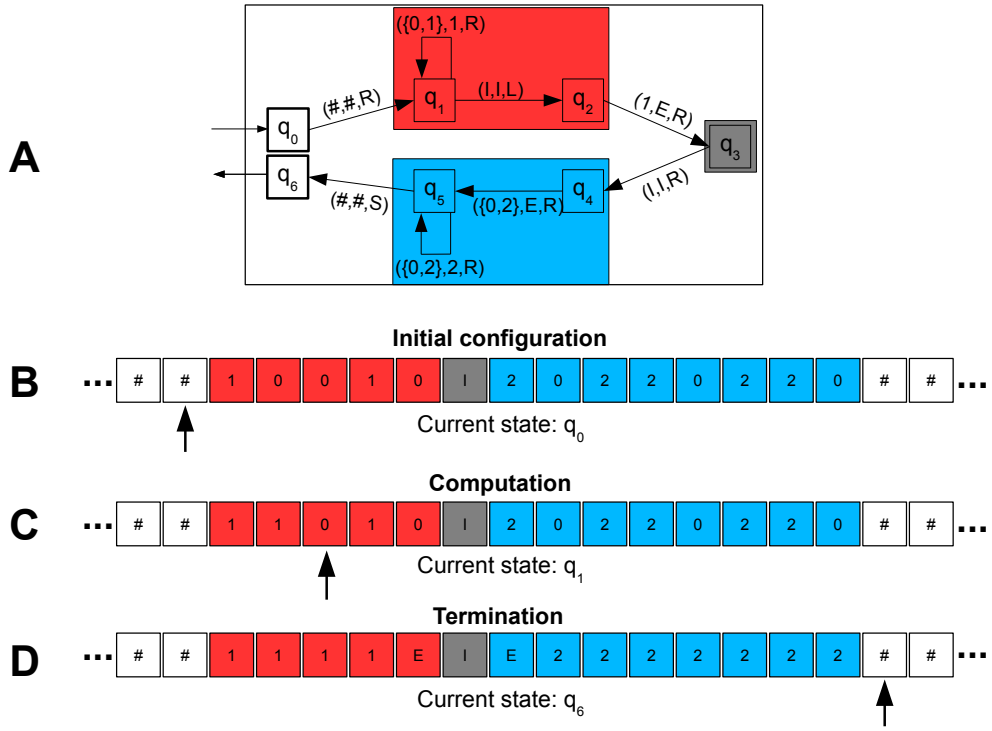


Figure A.1: *Example of a deterministic TM. The aim of the TM is to parse a string that consists of two domains separated by a boundary element (I, insulator). Domain 1 left to I is predominantly in state 1, whereas domain 2 is right to it and predominantly in state 2. The TM replaces all 0 symbols with their respective symbols from the domain they are located in. The boundary elements of the domains (i.e., the two cells adjacent to I) are furthermore marked by a special symbol E (end of domain). Black arrows mark the current position of the head of the TM.*
A: Finite state representation. Each rectangle represents a state of the TM, and the directionality of a transition is shown by an arrow. Each transition is labeled: the first element specifies the scanned symbol or symbols (denoted in curly brackets) that causes a particular transition when the TM is in the state the arrow originates from, the second one the symbol that is written, and the third one the direction in which the head moves. The TM is therefore specified as follows: $Q = \{q_0, \ldots, q_6\}$, $\Gamma = \{\#, 0, 1, 2, E, I\}$, *# as blank symbol,* $\Sigma = \{0, 1, 2, I\}$, $q_0 = q_0$, $F = \{q_6\}$, $\delta = \{q_0, \#\} \rightarrow \{q_1, \#, R\}$, $\{q_1, 1\} \rightarrow \{q_1, 1, R\}$, $\{q_1, 0\} \rightarrow \{q_1, 1, R\}$, $\{q_1, I\} \rightarrow \{q_2, I, L\}$, $\{q_2, 1\} \rightarrow \{q_3, E, R\}$, $\{q_3, I\} \rightarrow \{q_4, I, R\}$, $\{q_4, 0\} \rightarrow \{q_5, E, R\}$, $\{q_4, 2\} \rightarrow \{q_5, E, R\}$, $\{q_5, 0\} \rightarrow \{q_5, 2, R\}$, $\{q_5, 2\} \rightarrow \{q_5, 2, R\}$, $\{q_5, \#\} \rightarrow \{q_6, \#, S\}$.
B: Start configuration.
C: Snapshot during execution. The TM is in state q_1 *and currently parsing a cell with the tape symbol 0, which will be overwritten with 1 upon execution of the transition function.*
D: End configuration after the TM reached the final state q_6.

after the performed a particular reaction). I now provide a brief summary of how a TM operates (see also Figure A.1).

Initial configuration

A TM operates on a tape on which each cell i contains a particular symbol $\gamma_i \in \Gamma$. The tape is infinitely long either on the left side, the right side, or both, depending on the definition. For our purposes, this is irrelevant. A finite number of cells may contain valid input symbols $\sigma \in \Sigma$,

whereas the remaining cells are blank (# is therefore the only symbol allowed to be represented infinitely many times on the tape). The head is positioned at a particular start cell and the TM is in the designated start state q_0.

Computation

The TM then operates step-wise. In each step, it is in a particular state $q \in Q$ and the head is at a particular cell i that contains a tape symbol $\gamma \in \Gamma$. Unless $q \in F$, the TM operates as follows. It determines its new configuration from the current configuration by selecting and executing one or more "rules". A deterministic TM has at most one rule that is applicable for any given state and tape symbol, whereas for a non-deterministic TM, more than one rule is applicable for at least one given state and tape symbol. If multiple rules match, one is selected according to some arbitrary criterion (or randomly). Therefore, different runs of a non-deterministic TMs on the same input string may produce different results. A rule is applicable if its left side matches; that is, if the current state $q \in Q$ and the tape symbol $\gamma \in \Gamma$ at the cell where the head is positioned are as specified in the left side of the rule. After execution of the rule, the tape symbol may be changed and the head may change its position to adjacent cells (as specified by $\{\mathtt{L}, \mathtt{S}, \mathtt{R}\}$), where `L` denotes a movement to the left cell, `R` to the right cell, and `S` no movement — *stay*).

Termination

If the current state $q_c \in F$ (i.e., one of the final states), the TM halts and computation stops.

A.3 Mapping from a Turing Machine to the Chromatin Computer

In Section 3.2.4.2, I described how any TM can be mapped to a CC, using the notation of Bryant [13]. In this section, I want to give a specific example of such a mapping, based on the TM from Figure A.1. In Table A.2, I construct the mapping and list the corresponding CC rules for each transition from the TM as defined in Figure A.1. For the mapping, I utilize a $(6, 4, 2)$-CC (6 possible states, 4 positions per nucleosome, rules depending on 2 adjacent nucleosomes), as defined in Section 3.2.4. Each TM rule is encoded by two rules in the CC, which map to the binding and dissociation of the corresponding enzymes, respectively. Because the head of the TM is always at only one particular location, I note again that position 1 of each nucleosome is empty for all nucleosomes but one. The same applies to the current state of the TM, which is also stored at the nucleosome where the head is located. For more details, see Section 3.2.4.

Transition in TM	Corresponding transitions in the CC
$\{q_0, \#\} \to \{q_1, \#, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_0\#0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_0\#1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_0\#1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\#0]\ [H\ q_1\ -\ 0]$
$\{q_1, 1\} \to \{q_1, 1, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_1\ 1\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_1\ 1\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_1\ 1\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ 1\ 0]\ [H\ q_1\ -\ 0]$
$\{q_1, 0\} \to \{q_1, 1, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_1\ 0\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_1\ 0\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_1\ 0\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ 1\ 0]\ [H\ q_1\ -\ 0]$
$\{q_1, I\} \to \{q_2, \mathtt{I}, \mathtt{L}\}$	1: $[B\ B\ *\ 0]\ [H\ q_1\ I\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_1\ 0\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_1\ I\ 1]\ [B\ B\ *\ 0] \to [H\ q_2\ -\ 0]\ [B\ B\ I\ 0]\ [B\ B\ -\ 0]$
$\{q_2, 1\} \to \{q_3, \mathtt{E}, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_2\ 1\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_2\ 1\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_2\ 1\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ E\ 0]\ [H\ q_3\ -\ 0]$
$\{q_3, I\} \to \{q_4, \mathtt{I}, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_3\ I\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_3\ I\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_3\ I\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ I\ 0]\ [H\ q_4\ -\ 0]$
$\{q_4, 0\} \to \{q_5, \mathtt{E}, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_4\ 0\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_4\ 0\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_4\ 0\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ E\ 0]\ [H\ q_5\ -\ 0]$
$\{q_4, 2\} \to \{q_5, \mathtt{E}, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_4\ 2\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_4\ 2\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_4\ 2\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ E\ 0]\ [H\ q_5\ -\ 0]$
$\{q_5, 0\} \to \{q_5, 2, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_5\ 0\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_5\ 0\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_5\ 0\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ 2\ 0]\ [H\ q_5\ -\ 0]$
$\{q_5, 2\} \to \{q_5, 2, \mathtt{R}\}$	1: $[B\ B\ *\ 0]\ [H\ q_5\ 2\ 0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_5\ 2\ 1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_5\ 2\ 1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [B\ B\ 2\ 0]\ [H\ q_5\ -\ 0]$
$\{q_5, \#\} \to \{q_6, \#, \mathtt{S}\}$	1: $[B\ B\ *\ 0]\ [H\ q_5\#0]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_5\#1]\ [B\ B\ -\ 0]$ 2: $[B\ B\ *\ 0]\ [H\ q_5\#1]\ [B\ B\ *\ 0] \to [B\ B\ -\ 0]\ [H\ q_6\#1]\ [B\ B\ -\ 0]$

Table A.2: *Mapping from a TM to a chromatin computer and example of a specific chromatin computer program. The mapping from the TM as defined in Figure A.1 to a $(6, 4, 2)$-CC (6 possible states, 4 positions per nucleosome, rules depending on 2 adjacent nucleosomes) is shown (see Section 3.2.4). The left column lists the transitions of the TM from Figure A.1 and the right column gives the corresponding rules set for the CC. The notation is analogous to Bryant [13] (Figure 3.6), with one modification: for clarity, I enclosed each nucleosome by $[\ldots]$. B: special blank symbol that represents the absence of any chromatin mark; $*$: special symbol that is used to read any mark; $-$: indicates that the original symbol remained unchanged (only used for positions that are not uniquely determined); H: special symbol that stands for "head" and indicates the current position of the head of the corresponding TM for which the reversible mapping is constructed. For more details, see text, Section 3.2.4.1 and Section 3.2.4.2.*

Appendix B

Methodological Details for the Custom Array Design Pipeline, CEM-Designer Web Server, nONCOchip 2.0 and Alzheimer Custom Array

B.1 Methodological Details for the Custom Array Design Pipeline and the CEM-Designer Web Server

The CAD pipeline consists of a collection of *Perl* and *R* scripts and can be executed from the command line. It is not publicly available; however, it is used in the CEM-Designer web server and available upon request. The CAD pipeline requires the installation of several freely available programs: R [489], BLAT [381], liftOver, and twoBitToFa[1].

The CEM-Designer web server is implemented in Python and uses extensive JavaScript (jQuery in particular[2]) and Ajax for its dynamic interface. Therefore, JavaScript must strictly be enabled. The job and scheduling system of the CEM-Designer are based on the MITOS web server [490]. Currently, only one client executes jobs but this can be adjusted whenever more computational resources are needed.

An overview of all relevant steps of the CAD pipeline and the CEM-Designer web server are presented in Figure B.1 and Figure B.2.

[1] The latter three programs are available at http://genome.ucsc.edu, last accessed in August 2013.
[2] http://jquery.com/, last accessed in August 2013

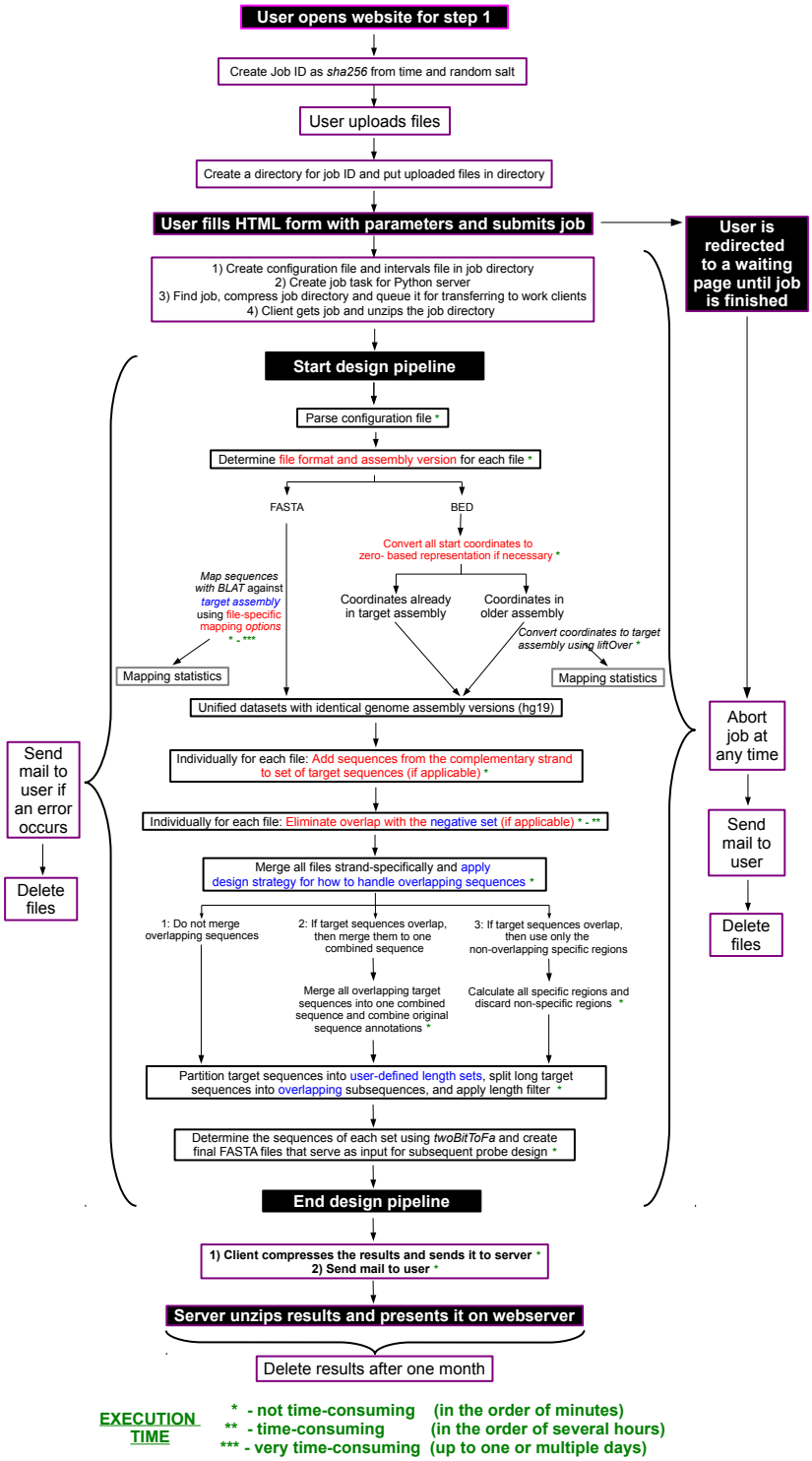


Figure B.1: *Summary of relevant steps of the CAD pipeline in conjunction with the CEM-Designer web server for step 1. Functions specific to the CEM-Designer web server are marked with a purple border. All customizable parameters appear at the corresponding step where they are applied (blue: general design parameters, red: file-specific parameters). The typical execution time for all relevant steps is also estimated (see legend).*

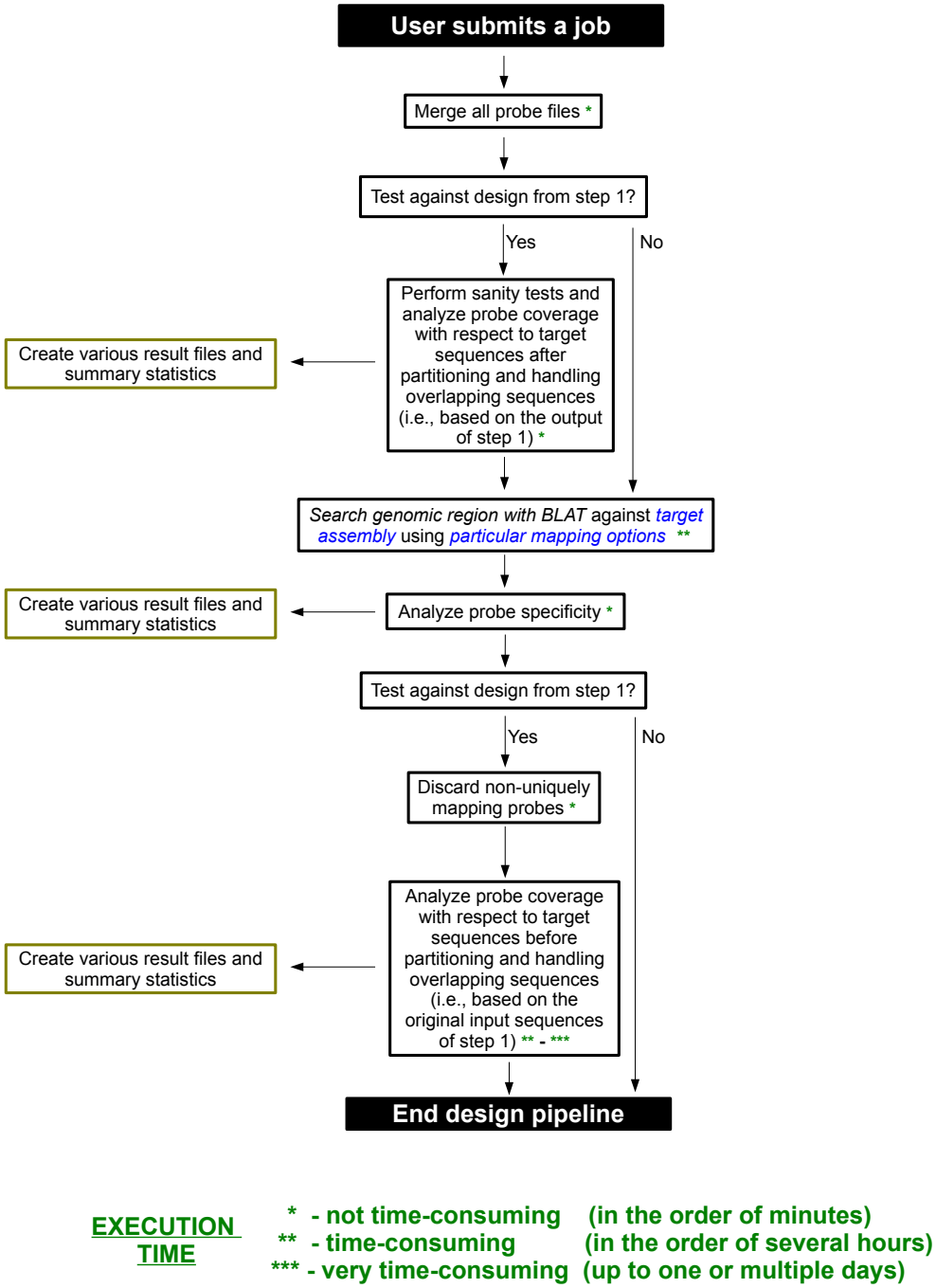


Figure B.2: *Summary of relevant steps of the CAD pipeline for step 3. Functions specific to the CEM-Designer web server have been omitted and are analogous to Figure B.1. All customizable parameters appear at the corresponding step where they are applied and are colored in blue, whereas output files are colored with an ocher boundary. The typical execution time for all relevant steps is also estimated (see legend).*

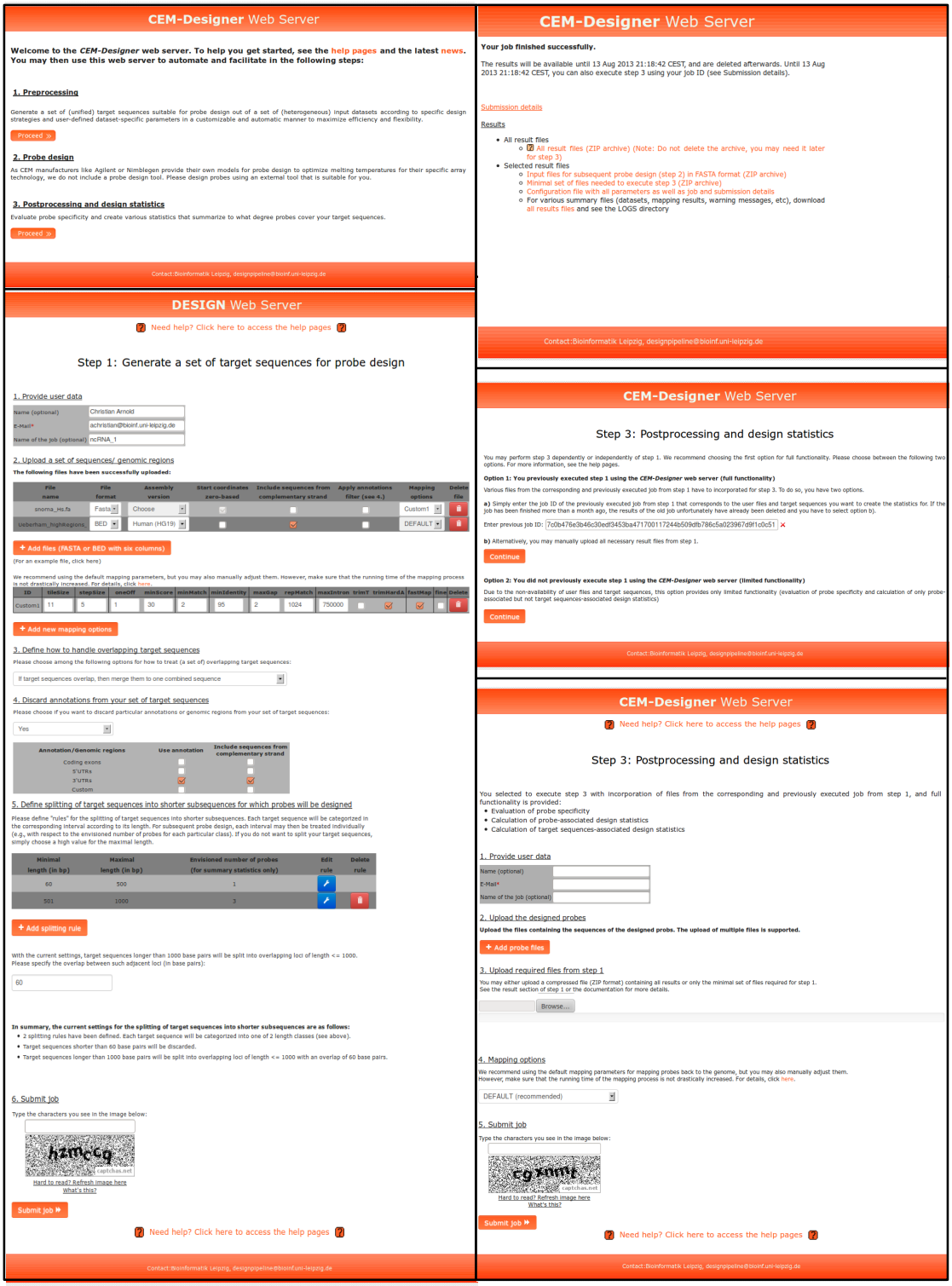


Figure B.3: *Snapshots of the CEM-Designer web server.*

B.2 Composition of the nONCOchip 2.0 and the Alzheimer Custom Array

B.2.1 nONCOchip 2.0

The nONCOchip 2.0 has been developed at the Fraunhofer Institute for Cell Therapy and Immunology (Fraunhofer IZI, Leipzig, Germany) to facilitate the analysis of expression changes of disease-associated lncRNAs in parallel with protein-coding genes (human genome version hg19). It comprises 913,413 probes of 60 bp each. The design strategy ensures that all probes are unique on the DNA level (according to hg19) as well as on the RNA level (according to *RefSeq*). Overall, 65% of probes correspond to lncRNAs and the remaining probes to protein-coding mRNAs (20%) and regions antisense to 5' or 3' UTRs (11%) (Tables B.1 and B.2). In addition to *Agilent*'s 026652 probeset (composed of *RefSeq* Build 36.3, *Ensembl* Release 52, *Unigene* Build 216, and *GenBank* from April 2009), I designed 190,362 (20.4%) probes for all protein-coding genes as annotated in *Gencode* v4 such that a total of 93% of all protein-coding genes and 95% of their corresponding alternative transcripts are represented by at least one probe on the microarray. Noteworthy, the nONCOchip 2.0 contains more than 160,000 probes (18%) for ncRNAs that are not available on any other commercial microarray platform. These probes represent RNAs that have been found to be controlled by major cancer-related pathways (e.g., the oncogene *Stat3*, the tumor suppressor protein *TP53*, or cyclins) and have been detected by transcriptome-wide expression variation studies utilizing *Affymetrix* Human Tiling Arrays also performed at the *Fraunhofer Institute for Cell Therapy and Immunology* in cooperation with the University of Leipzig (tiling array regions). They are largely undescribed so far[3]. Furthermore, the nONCOchip 2.0 contains a comprehensive representation of known lncRNAs retrieved from public databases (*NONCODE* [491], *lncRNAdb* [492], *fRNAdb* [493], *RNAdb* [494], *H-InvDB* [495], *Gencode* v4 [380], *RefSeq*[496], literature — lncRNAs originating from actively transcribed genes [201], chromatin-associated RNA [188], snoRNAs from the snoBoard database [497], intronic RNAs as identified by Nakaya et al. [429], in total 42% of probes) as well as long lncRNAs with conserved secondary structure (RNAz [384], Evofold [382], 8%). Since natural antisense transcripts appear to regulate transcription and translation of neighboring genes [498], I designed 156,698 (16.8%) probes mapping antisense to protein-coding genes (*Gencode* v4). Lastly, nONCOchip 2.0 contains probes from the nONCOchip 1.0 that do not overlap with probes from the nONCOchip 2.0.

For each dataset, I also calculated various statistics to analyze to what extend the intended design strategies worked as expected (e.g., histograms of the number of probes per target sequence, target sequence length or the correlation of target sequence length and the number of fully overlapping probes) (Figure B.4).

[3]Hackermüller et al. "Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro ncRNAs", submitted.

Table B.1: *Overview of probe distribution for the nONCOchip 2.0. A probe corresponds to a category if it overlaps to at least 95% (57 nucleotides) with at least one annotation of the category. Because categories may overlap, a probe may be associated with multiple categories. The relative fraction is defined according to overall number of probes on the nONCOchip 2.0, and the numbers may not add up to 100% due to the mandatory control probes.*

Category	Number of probes	Relative fraction (in %)
Transcriptome-wide studies	160,884	17.61
-Cell cycle	60,178	6.59
-TP53	81,455	8.92
-Stat3	30,372	3.33
Predicted ncRNAs	77,771	8.51
ncRNAs from public databases	380,608	41.67
ncRNAs, in total	598,428	65.52
Protein-coding mRNAs (*Gencode* v4 annotation)	180,824	19.80
Probes on the reverse complementary strand of UTRs (5' or 3')	99,387	10.88

Table B.2: *Genomic distribution of probes for the nONCOchip 2.0. A probe corresponds to a category if it overlaps strand-specifically to at least 95% (57 nucleotides) with at least one annotation (i.e., feature or sequence) of the category. For introns and intergenic regions, the strand information is ignored. 5'UTRs and 3'UTRs correspond to 5' and 3' untranslated regions of mRNAs. CDS corresponds to the coding exons of mRNAs. The relative fraction is defined according to overall number of probes on the nONCOchip 2.0. Similar to Figure B.1, the numbers may not add up to 100% due to the mandatory control probes and probes that overlap with no category with at least 95%, for example.*

Annotation category	Number of probes	Relative fraction (in %)
5'UTRs	38,249	4.19
CDS	65,165	7.13
3' UTRs	91,642	10.03
Introns	409,884	44.87
Intergenic regions	172,376	18.87

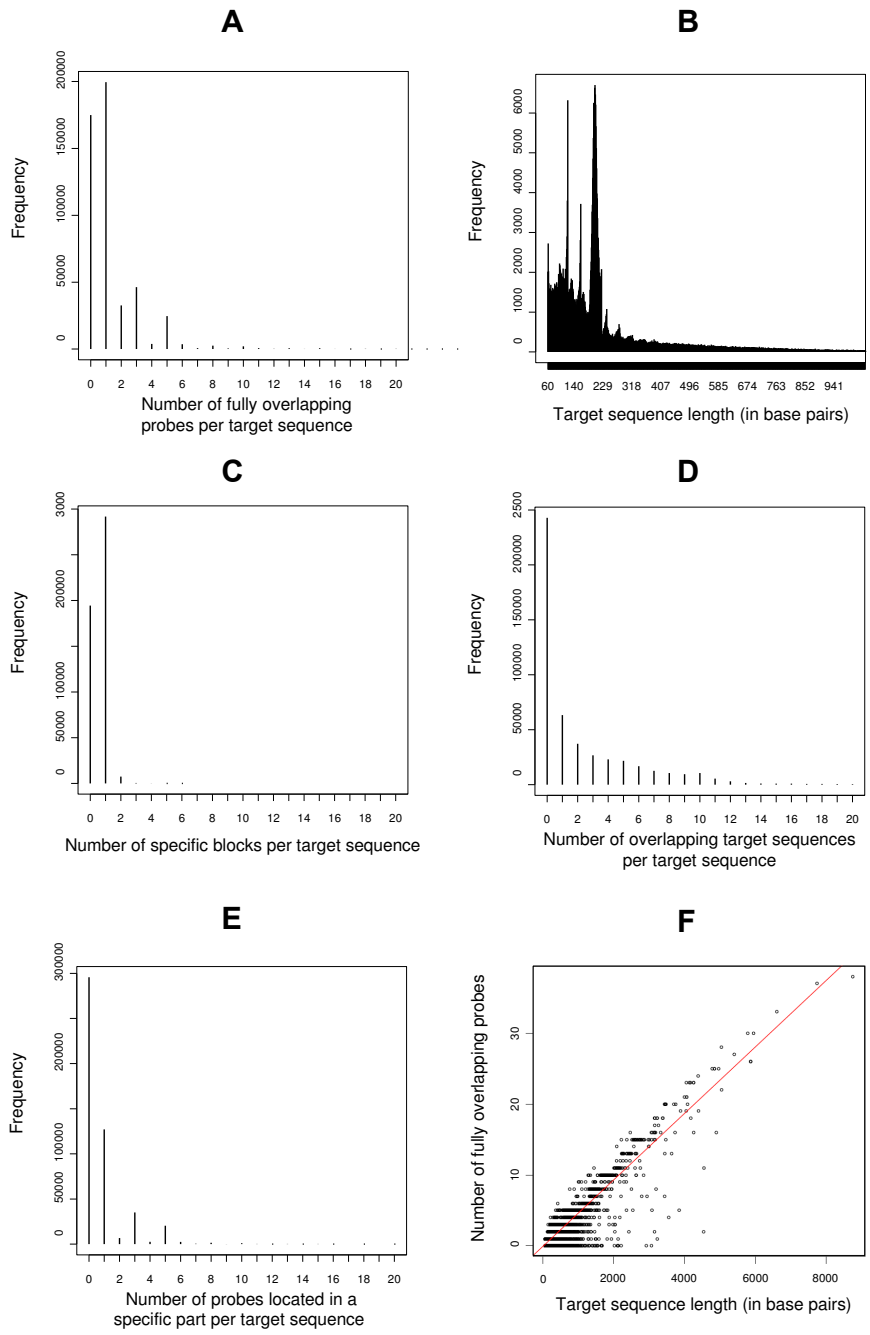


Figure B.4: *Design statistics for the collective ncRNA dataset from the* `nONCOchip` *2.0 after the first probe design. Various histograms (A–E) and one correlation plot (F) are shown. For clarity, only a fraction of the values on the x-axis are shown (maximum values: a:1,457; b:415,001; c:136; d:1,340; e:458). In F, a random sample of 20,000 target sequences has been selected for the correlation plot.*

For both `nONCOchip` 2.0 and the `Alzheimer Custom Array` (see Section B.2.2), *Agilent SurePrint*

Table B.3: *Genomic distribution of probes for the `Alzheimer Custom Array`. The table is analogously composed as Table B.2, and I refer to Table B.2 for details.*

Annotation category	Number of probes	Relative fraction (in %)
5'UTRs (sense)	39,233	4.21
5'UTRs (antisense)	38,021	4.08
CDS (sense)	70,451	7.56
CDS (antisense)	43,799	4.70
3'UTRs (sense)	101,297	10.87
3'UTRs (antisense)	73,340	7.87
Introns	388,881	41.73
Intergenic regions	162,803	17.47
Pseudogenes	8,201	0.88
Repeats	17,706	1.90

G3 Custom Gene Expression Exon microarrays[4] were used that comprise roughly one million features. The probes have a length of 60 bp and are much more sensitive to expression changes than shorter probes [e.g., see 499].

B.2.2 `Alzheimer Custom Array`

The `Alzheimer Custom Array` has been designed to facilitate the analysis of expression changes for a larger set of samples with reliable variance estimations for Alzheimer's disease (AD). In total, I had 19 AD patient samples and 22 control samples. The `Alzheimer Custom Array` is in large parts identically composed as the `nONCOchip` 2.0 (Table B.3) but differs in a few notable details. It does not contain probes from the `nONCOchip` 1.0 and excludes a few probes from tumor-relevant signaling pathways (see above) but instead includes almost 7,000 probes from AD-associated genes and a set of roughly 600 mRNAs, previously identified to contain single nucleotide polymorphisms related to AD [500] (compiled by Uwe Überham). Additionally, almost 200,000 probes for transcripts that were significantly highly expressed (FDR < 0.05) or significantly differentially expressed (FDR < 0.05) based on a preliminary study with one AD patient sample and one control brain tissue sample assessing unbiased transcriptome-wide expression variations utilizing the Affymetrix Human Tiling 1.0 array set are included.

Overall, 931,898 probes have been designed for the `Alzheimer Custom Array` (26,701 *Agilent* probes and 905,197 custom probes). 22.6% of all probes represent isoforms of protein-coding genes. Analogously to the `nONCOchip` 2.0, probes represent ncRNAs from public databases, genomic loci containing conserved secondary structures, novel transcripts controlled by major disease-relevant pathways, and novel transcripts identified to be associated with AD by a preliminary study (see

[4] Exon arrays contain probes that can be arbitrarily distributed along the entire gene, unlike standard expression arrays that contain probes biased toward the 3' end of genes.

above). This corresponds to 378,744 (40.6%), 72,877 (7.8%), 232,473 (24.9%) and 127,127 (13.6%) probes, respectively. lncRNAs [189, 201] are covered by 36,052 (3.8%) probes.

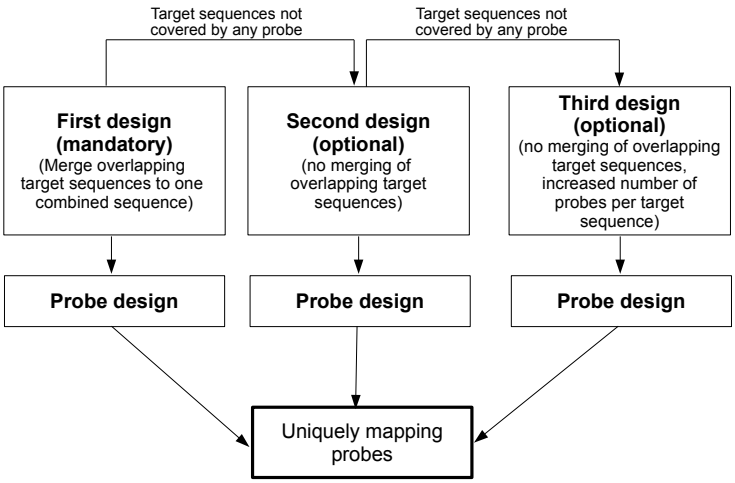
B.3 Methodological Details for the Application of the CAD Pipeline for the nONCOchip 2.0 and the Alzheimer Custom Array

I used the CAD pipeline for the preprocessing of target sequences for both the nONCOchip 2.0 and the Alzheimer Custom Array. If only sequences were known for a particular dataset and not their genomic positions, I used BLAT for the sequence mapping to the hg19 genome and considered only hits if they had at least 95% sequence identity. If multiple hits were found, I used only the best-scoring one. If hits spanned multiple blocks due to introns, I treated each block as a separate sequence. After unifying genomic coordinates among all datasets, eliminating duplicate sequences (i.e., sequences with identical start and end coordinates), and applying a strategy for how to handle overlapping sequences, I partitioned all sequences into three distinct sets and designed probes individually for each set (Figure B.5). Because *Agilent* uses 60-mer probes, I additionally eliminated target sequences shorter than 60 bp. Furthermore, for datasets containing ncRNAs, I eliminated any overlap with coding regions (according to *Gencode* v4) to avoid probe overlap. To represent each genomic loci only once, I merged overlapping target sequences to a combined sequence in the first design round (second strategy, see Section 5.2.1.3 and below).

For probe design, I used the eArray software from *Agilent* and did not specify any preferred probe position (i.e., probes were placed randomly across the sequence). Due to the impossibility of choosing the full hg19 genome as reference genome, I used the default reference transcriptome (which includes all *RefSeq* annotated transcripts) as provided by *eArray*. After probe design, I tested probe specificity by similarity search against the genome by mapping all probes to the hg19 genome using BLAT with settings that aim to improve sensitivity (`-stepSize=5 -repMatch=1000000 -fine -minIdentity=90`, whereas all other settings have default values). I then discarded any probes that mapped non-uniquely (two or more hits with at least 95% identity).

After the first design, I calculated various coverage statistics (such as histograms and correlation plots, see Figure B.4). In particular, I calculated what percentage of target sequence before application of the CAD pipeline are represented by at least one specific probe (hereafter denoted $c_{>0}$, see also Figure B.4 A). If $c_{>0}$ was smaller than 70%, I created a reduced dataset containing only those target sequences not yet represented by any specific probe and performed up to two additional designs rounds (including probe design using *eArray*, see above) with modified design parameters (see below and Figure B.5 B). In contrast to the first design, overlapping sequences were not merged in the second design, which expectedly produced overlapping probes but also substantially increased $c_{>0}$. In the third design, to further improve $c_{>0}$, I used the same parameters as for the second design but increased the number of probes per target sequence (Figure B.5 A).

A



B

Target sequence length	Probes per target sequence		
	First design	Second design	Third design
60-300 bp	1	1	3
301-600 bp	3	3	5
601-1000 bp	5	5	10
>1000 bp	Split into loci with length ≤1000 with overlap of 60 bp		

Figure B.5: *General design strategy for the nONCOchip 2.0 and the Alzheimer Custom Array.*
A: Visualization of the general strategy to maximize the number of target sequences that are represented by at least one uniquely-mapping probe. Up to three independent probe designs were made, each of which with different design parameters. Ultimately, uniquely mapping probes from all designs are merged. For more details, see text.
B: Summary of the number of probes per target sequence after the partitioning of target sequences into three distinct sets for the first, second, and third design.

Indeed, particularly for short target sequences (for which only one probe was designed in the first two designs), I often successfully designed a considerable amount of new specific probes (see Section 5.3 for a discussion). I then merged all specific probes from all designs and recalculated $c_{>0}$. Finally, I removed duplicate probes by identifying probes that overlap with at least 59 bp with other probes and then randomly deleted all but one of these overlapping probes.

Appendix C

Methodological Details and Additional Results for the `Alzheimer Custom Array`

C.1 Identification of Differentially Expressed Probes

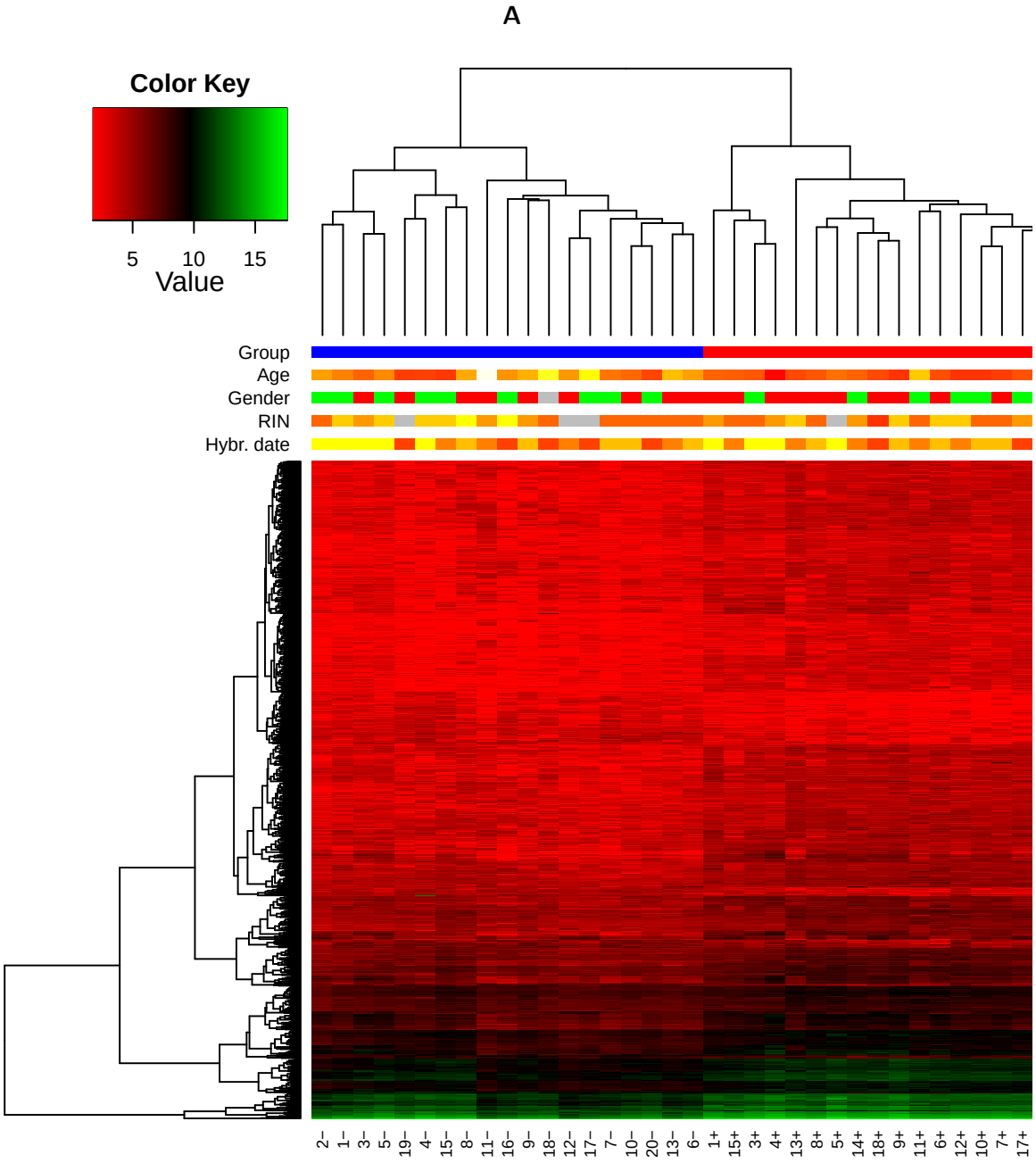
To identify differentially expressed probes, I defined the linear model:

$$E[X_i] = \alpha \times \mathsf{AD} + \beta \times \mathsf{Age} + \epsilon \tag{C.1}$$

, where $E[X_i]$ is the expected expression of probe i, ϵ an error term, α the coefficient modeling the impact of AD on the expression variance of probe i, and β the coefficient modeling the influence of the patient's age. I used the `Limma R` package [501] for model fitting, and reliable variance estimations were obtained by empirical Bayes moderated t-statistics.

The obtained p-values have to be adjusted for multiple testing because in the analysis of any high-dimensional data such as microarray data, massive parallel testing introduces a multiple testing problem. For this, I controlled the false discovery rate (expected proportion of Type I errors among all significant hypotheses). Because the Benjamini-Hochberg (BH) procedure [502] suffers from multiple flaws [503, 504], I employed a modified BH procedure that incorporates an estimated proportion of the null p-values [503] to compute q-values (false discovery rate analogue of the p-value, or the minimum false discovery rate at which a test may be called significant). This offers distinct advantages over the traditional BH method (see [504] for details) and is implemented in the `fdrtool R` package [504, 505]. A comparison between the traditional BH and the `fdrtool` adjustment revealed that for a given q-value, slightly more probes were deemed significant with `fdrtool` but all probes deemed significant using a BH adjustment were also significant with `fdrtool`.

All following analyses were based on the set s_{diff} of the 4,095 uniquely mapping, differentially expressed probes.



(Continued on next page.)

C.2 Identification of Differentially Expressed Loci

To identify differentially expressed loci, I integrated different annotation sources with known (e.g., *Gencode* v14, including *Gencode* v14 long non-coding RNAs, and a ncRNA list collected by Cabili et al. [58]; see Section B.2 for a full list) and unknown transcript structures and strand (loci from

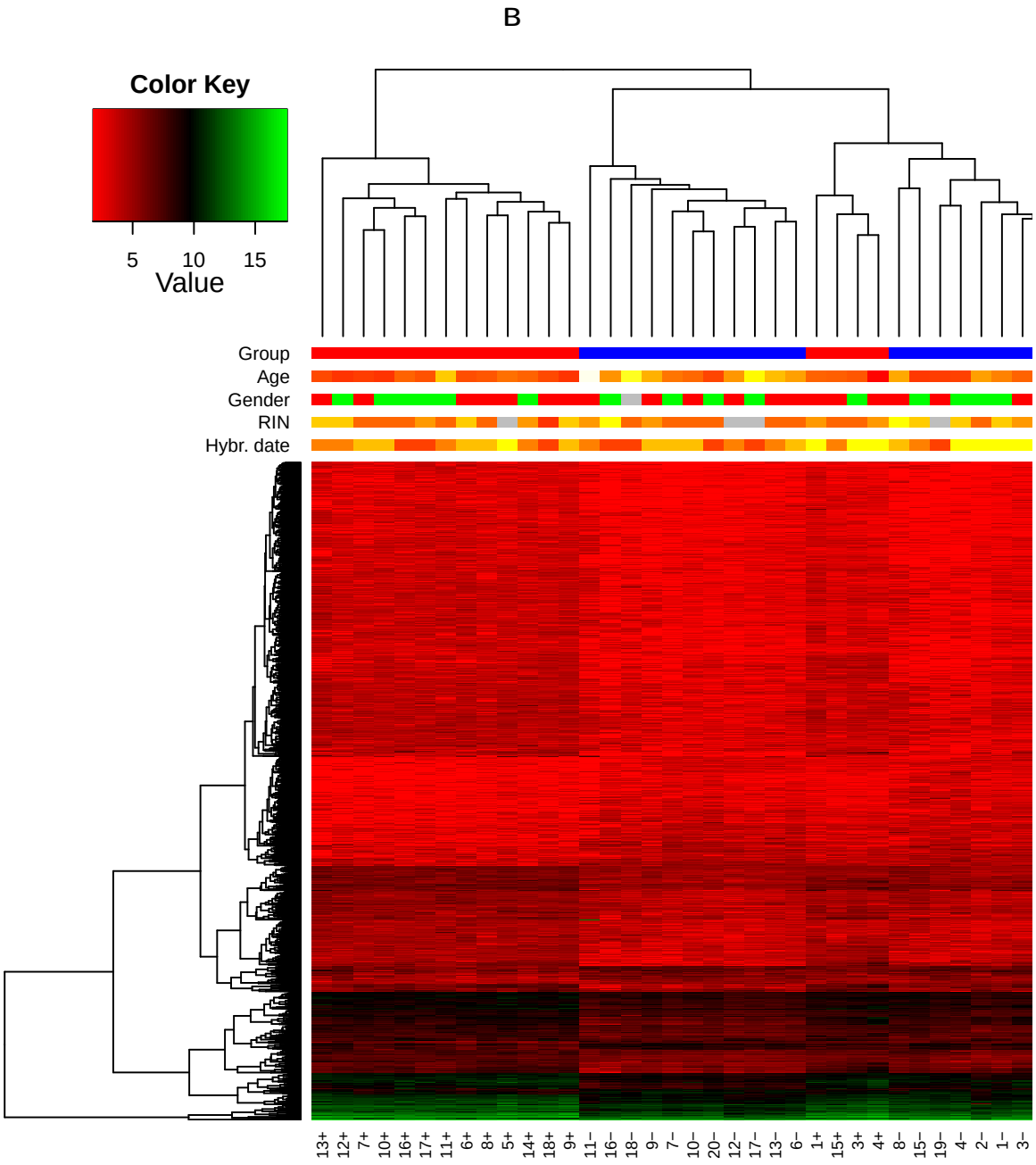


Figure C.1: *Example heatmaps of differentially expressed probes without row scaling. The figure is identical to Figure 6.3, except that values have not been centered and scaled in any direction.* ***A****: based on q=0.1.* ***B****: based on q=0.2. See Figure 6.3 for further details. (Continued from previous page.)*

the tiling array experiments, ncRNA predictions, and caRNAs [188]).

Because different annotations may overlap, I used *Gencode* v14 as primary annotation and complemented this with the manually collected annotations to maximize the information content. Analogous to the enrichment analysis, probes needed at least 95% overlap with a particular annotation item (e.g., exon or intron) to be considered.

For each probe $p \in s_{\text{diff}}$, I determined whether p is located in a locus with known transcripts (protein-coding, non-coding, or pseudogenes, as described before). Probes were mapped to a particular gene if it was located in (i) an exon of at least one annotated splicing variant (only for protein-coding transcripts because non-coding and pseudogene transcripts may exist in an unspliced and/or spliced version), (ii) the UTR of that gene, or (iii) in a putative previously unrecognized exon (no overlap with annotated exons but located in an exon of at least two spliced ESTs). Probes located exclusively intronic of a protein-coding gene (i.e., no overlap with annotated exons and less than two overlaps with exonic ESTs) were classified as putative intronic transcripts and therefore added to the non-coding list. If multiple introns overlapped, I used the cluster of overlapping introns as loci.

Importantly, I considered only sense and discarded antisense overlap because the transcript structure is not known for transcripts that are antisense to annotated transcripts unless they map to known antisense transcripts, which were already included in the various annotation sources as listed above. However, for annotation items with an unknown reading direction (e.g., loci from the tiling array experiments, ncRNA predictions, caRNAs [188]), I ignored the strand information and considered all overlaps in sense direction to not misleadingly lose relevant hits.

I tested the full annotated gene associated with a particular probe $p \in s_{\text{diff}}$ for significance rather than only the overlapping transcript(s) itself, due to the difficulty and uncertainty of distinguishing among individual splicing variants. If a probe mapped to multiple distinct annotated genes, I tested each gene individually but recorded the ambiguity in order to not lose any potentially relevant signals.

For each differentially expressed probe $p_i \in s_{\text{diff}}$ that overlapped with a particular differential expression candidate i (i.e., a locus with known or unknown transcript structure) in sense direction, I then identified the set of probes p_{all_i} that also overlapped with i with the criteria as described above (with respect to their genomic location such as exonic or intronic) and recorded the fraction of probes for which the expression level change was in the same direction as p_i (i.e., up- or downregulated as compared to the control group). I reasoned that if a particular gene is differentially expressed, the probes that map to this gene (regardless of whether they are deemed as significantly expressed) should at least have an expression level change in the same direction as the differentially expressed probe. I then used a one-tailed binomial test to identify differentially expressed loci with a significance threshold of $p = 0.05$. Because significance can only be achieved with a set of at least

five other probes, I separately recorded cases when the test could not become significant because not enough probes have been designed and when more than 50% of the overlapping probes had a expression level change in the same direction (1 out of 1, 2 out of 3, 3 out of 4, ...). Additionally, I recorded transcripts that achieved borderline significance (4 out of 5, 5 out of 6, and 6 out of 7). These loci should be treated with caution, however, because they may contain an increased amount of false positives.

For probes located in loci with unknown transcript structure(s), I checked if the probe overlaps with spliced ESTs. If more than one spliced EST overlapped with the probe, I used the full overlapping EST cluster as locus rather than the original locus for the subsequent significance test.

Lastly, for each of the four classes (three types of known transcripts and unknown transcripts), I filtered the list and only retained loci for which either the binomial test was significant or for which at least one probe had a q-value smaller than 0.05. Although this strict procedure may eliminate potentially relevant signals, it reduces the number of false positives due to the relatively high initial q-value.

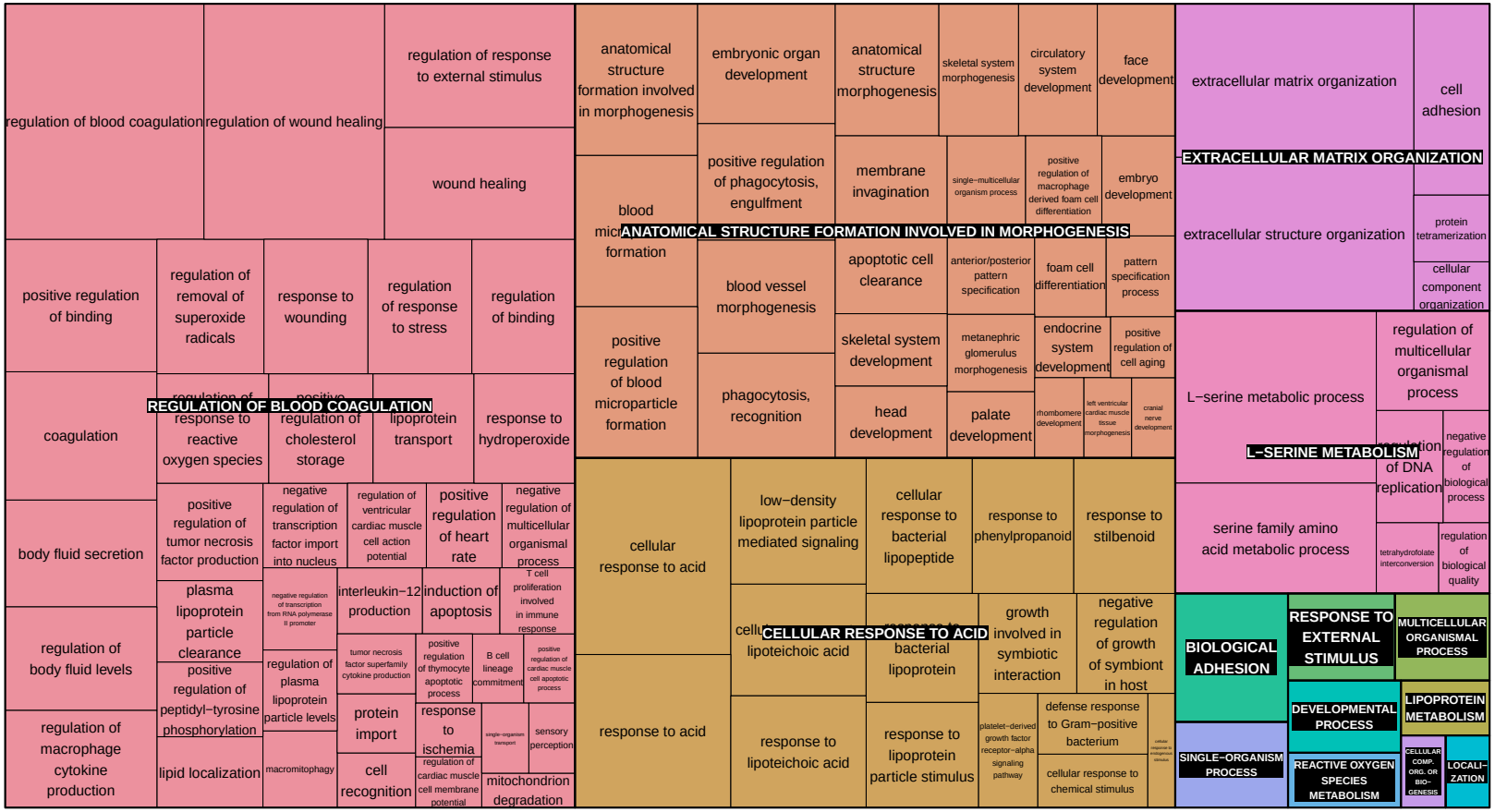
C.3 Functional Characterization of Differentially Expressed Loci and Overlap with Known AD-Associated Loci

C.3.1 Functional Characterization of Differentially Expressed Loci

In the `GO-TermFinder` web interface, I did not exclude any `GO` evidence codes and selected the *UniProt-GOA* gene association file. Because the GOA group annotates to the primary protein accession instead of the gene accession, individual Ensembl IDs may have multiple protein accessions, and `GO-TermFinder` ignores them due to the ambiguity. Because this affected a substantial amount of Ensembl IDs and therefore unnecessarily reduced the size of the dataset (both in the background list and the list of differentially expressed loci), as a workaround (as suggested by John Matese[1]), I first manually converted Ensemble IDs into *UniProtKB* accession numbers using the *UniProt* ID Mapping tool (`http://www.uniprot.org/`, last accessed in August 2013). Because I identified full genes in the differential expression analysis rather than particular isoforms, I retained all *UniProtKB* accession numbers if a particular Ensembl ID mapped to multiple *UniProt* IDs. For differentially expressed pseudogenes, only 2 out of 29 could be mapped, and an enrichment analysis was therefore not possible.

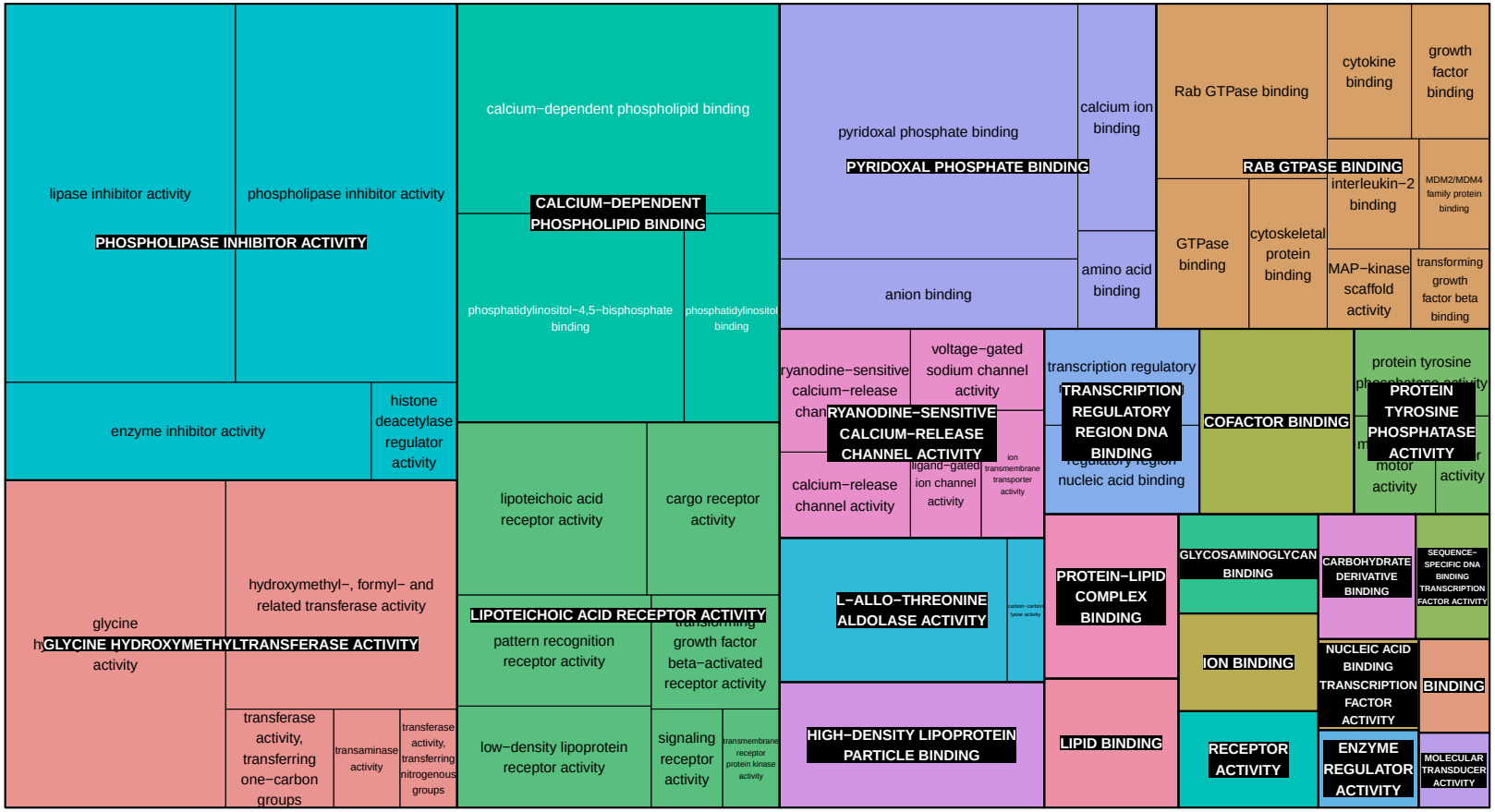
[1] Director of the Princeton University *MicroArray database* at the *Lewis-Sigler Institute for Integrative Genomics*

A



(Continued on next page.)

B



(Continued on next page.)

C

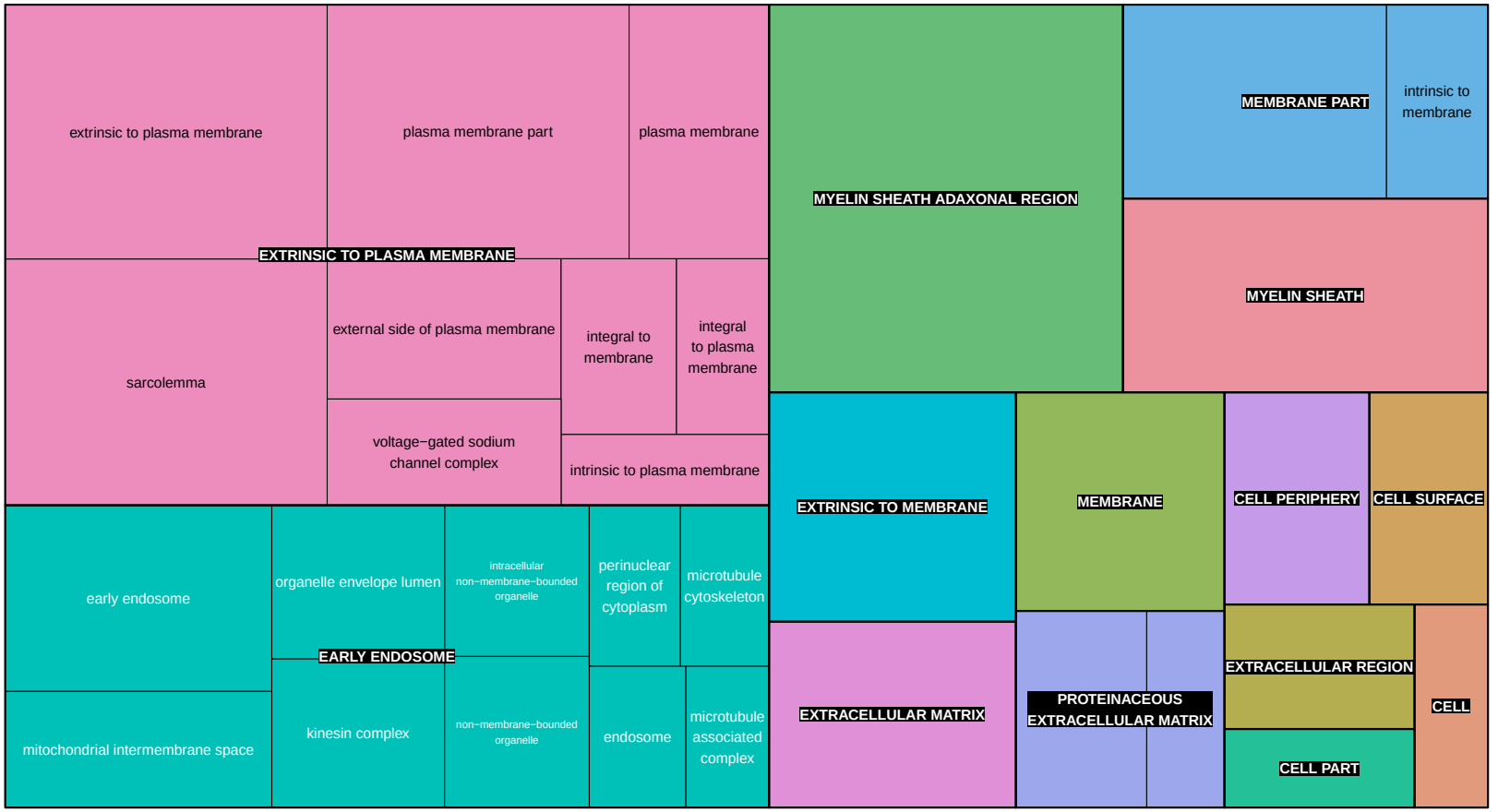
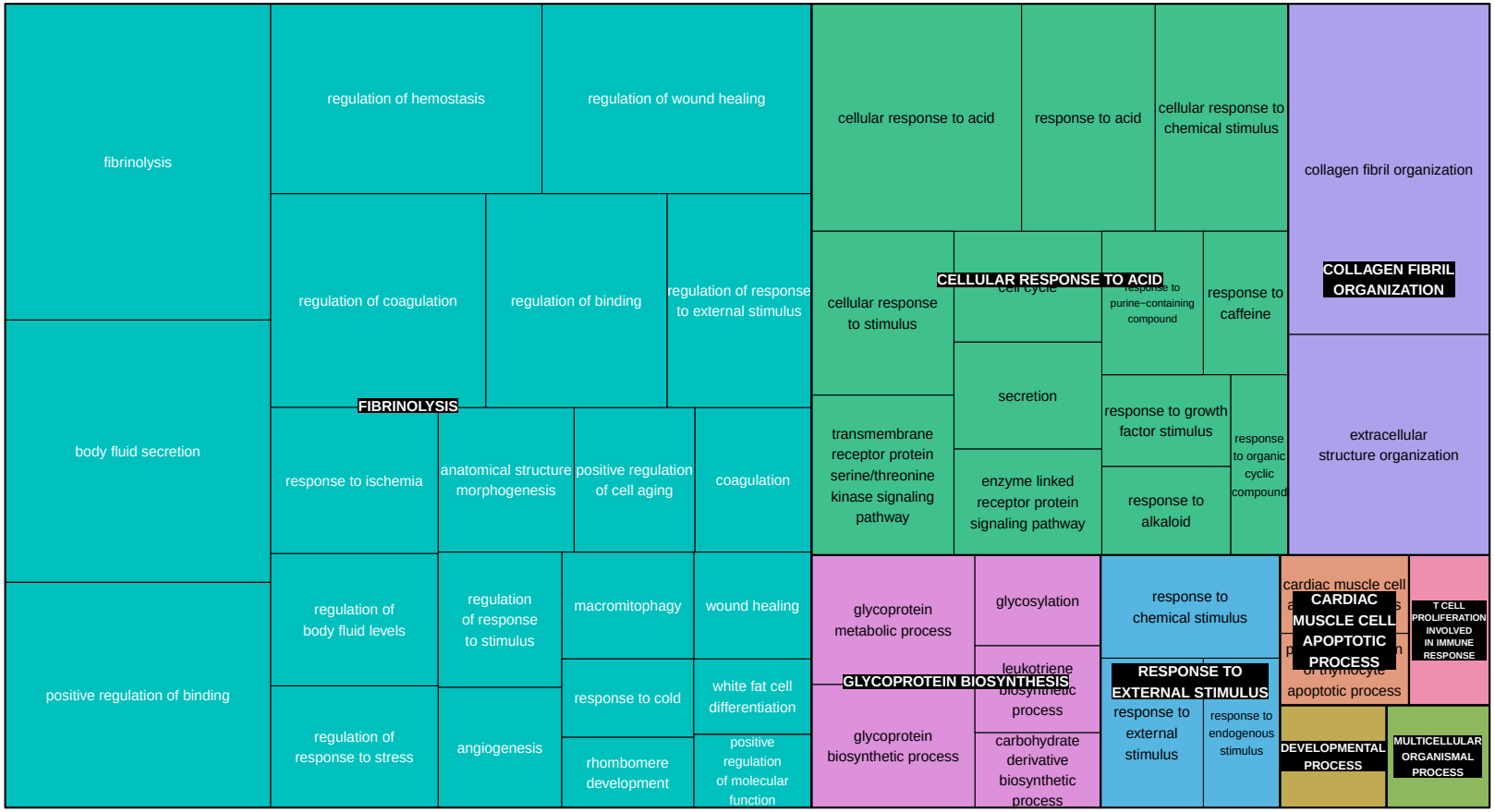


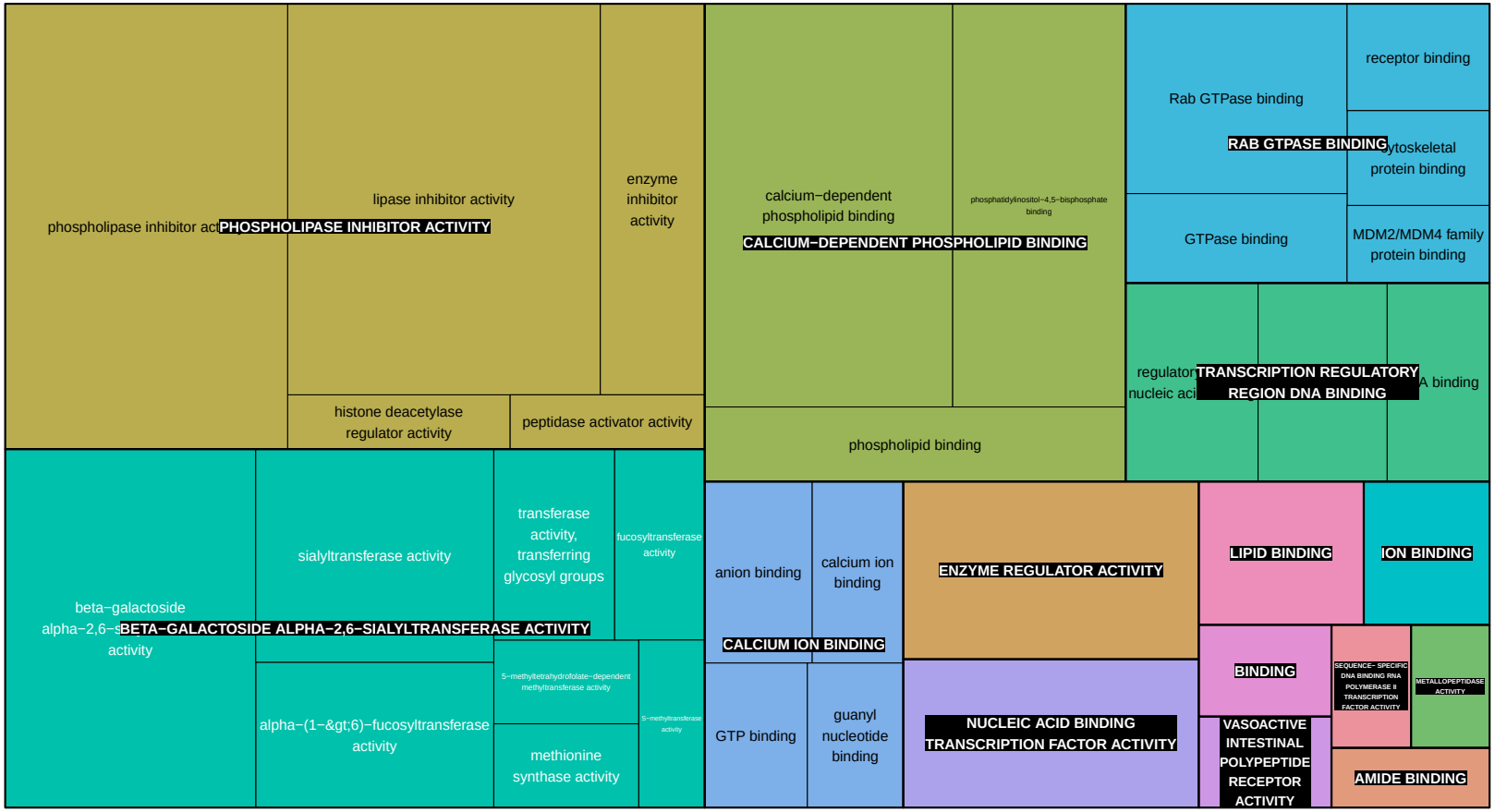
Figure C.2: *Results of the GO terms enrichment analysis for putative differentially expressed protein-coding genes.* ***A****: Biological process,* ***B****: Molecular function,* ***C****: Cellular component. See text for further details. (Continued from previous pages.)*

A



(Continued on next page.)

B



(Continued on next page.)

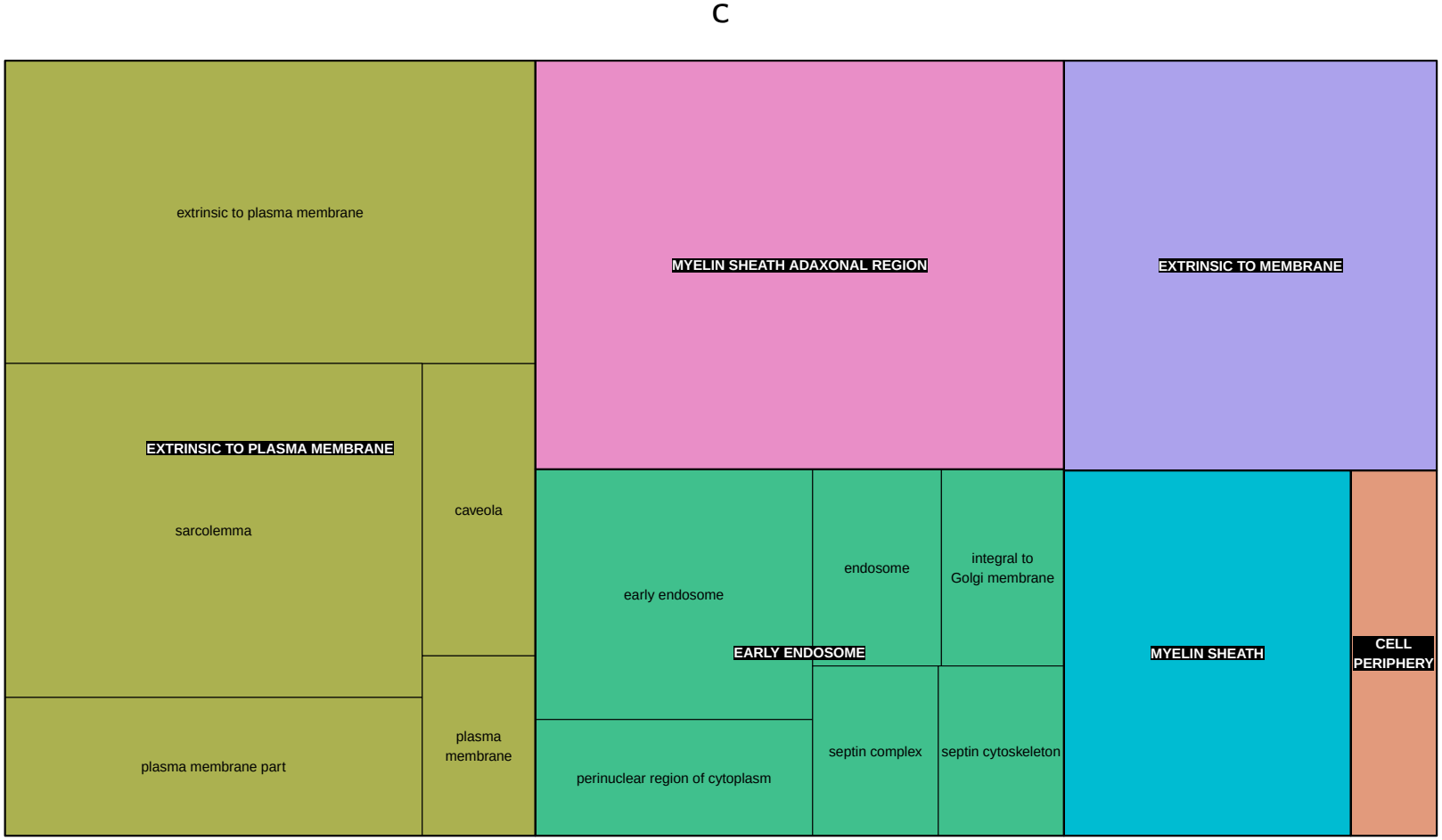


Figure C.3: *Results of the GO terms enrichment analysis for putative differentially expressed non-coding genes and transcripts. For non-coding loci originating from introns, I selected the surrounding gene for the GO terms enrichment analysis (if available).* ***A****: Biological process,* ***B****: Molecular function,* ***C****: Cellular component. See text for further details. (Continued from previous pages.)*

C.3.2 Overlap with Known AD-Associated Loci

Table C.1 lists differentially expressed loci that overlap either in sense or antisense direction with known AD-associated genes according to a manually created list collected by Uwe Überham, based on publications of Bossers et al. [476], Bertram et al. [506], Tan et al. [507], Ravetti et al. [508], Colangelo et al. [509], and Ginsberg et al. [510].

C.4 Splice Site Conservation Analysis

The protein-coding background set consisted of *RefSeq* annotated coding transcripts (∼120,000 transcripts with ∼347,000 unique splice sites) and the non-coding background of the *Gencode* v14 long non-coding RNA set (∼21,300 transcripts with ∼63,000 unique splice sites).

Each splice map was based on the 46-way `multiz` alignment from the *UCSC* Genome Browser. It contained the exact aligned genomic coordinates for all available species and the corresponding `MaxEntScan` score for each listed splice site. The `MaxEntScan` can be considered as a similarity measure for a splice site motif, and scores > 3 strongly indicate strongly that a splice site is functional [115, 511]. We considered an aligned splice site in a particular species as functional if either annotated by *RefSeq* or an EST or if *MaxEntScan* > 3.

We assessed the significance of the observed values as follows. First, we drew 1,000 random samples from the corresponding background set and randomly selected the same number of genes as in the set of differentially expressed genes. The significance of the observed signal was then calculated as the fraction of the random background samples that had a value equal or higher (if the signal had a value higher than the background) and equal and lower (if the signal had a value lower than the background) than the signal, respectively. I used a threshold of 0.05 for significance.

Table C.1: *Overlap with known AD-associated genes. The table lists all differentially expressed loci that overlap either in sense or antisense direction with known AD-associated genes according to various references.* ***AD gene****: Gene name with reported AD association as reported by the reference.* ***Ensembl ID****: Ensembl ID of the AD gene.* ***Signal origin****: Characterization of the location from which the signal (differentially expressed probe) originates from.* ***Reg. in AD****: Is the transcript downregulated (−) or upregulated (+) in AD in the* `Alzheimer Custom Array`*?* ***Overlap direction****: Direction of overlap with the reported gene from the literature. For more details, see text.*

AD gene	Ensembl ID	Signal origin	Reg. in AD	Overlap direction	Reference
APOA2	ENSG00000158874	exon of protein-coding gene	+	sense	[506]
MME	ENSG00000196549	intron of protein-coding gene	+	sense	[506]
MYOZ3	ENSG00000164591	exon of protein-coding gene	+	sense	[476]
GRIA1	ENSG00000155511	intron of protein-coding gene	−	sense	[476]
NEDD9	ENSG00000111859	intron of protein-coding gene	−	sense	[510]
ADCYAP1R1	ENSG00000078549	intron of protein-coding gene	+	sense	[506]
CAV1	ENSG00000105974	intron of protein-coding gene	+	sense	[506]
NRG1	ENSG00000157168	intron of protein-coding gene	+	sense	[506]
SLC18A3	ENSG00000187714	exon of protein-coding gene	+	sense	[506]
CALHM1	ENSG00000185933	exon of protein-coding gene	+	sense	[506]
TCF7L2	ENSG00000148737	intron of protein-coding gene	+	sense	[506]
ADAM12	ENSG00000148848	intron of protein-coding gene	+	sense	[506]
BDNF	ENSG00000176697	exon of protein-coding gene	−	sense	[506], [509]
ATXN8OS	ENSG00000230223	antisense lncRNA	+	sense	[506]
SAMD4A	ENSG00000020577	intron of protein-coding gene	−	sense	[507]
FOS	ENSG00000170345	intron of protein-coding gene	+	sense	[510]
TGFB3	ENSG00000119699	exon of protein-coding gene	−	sense	[507]
LIPC	ENSG00000166035	intron of protein-coding gene	+	sense	[506]
IGF1R	ENSG00000140443	intron of protein-coding gene	+	sense	[506]
TP53	ENSG00000141510	exon and intron of protein-coding gene	+	sense	[506]
RUNX1	ENSG00000159216	intron of protein-coding gene	+	sense	[506]
RBM3	ENSG00000102317	exon of protein-coding gene	−	sense	[507]
STARD7	ENSG00000084090	intron of protein-coding gene	+	antisense	[476]
TTN	ENSG00000237298	antisense lncRNA	−	antisense	[508]
EFNA5	ENSG00000184349	RNAz prediction, intronic ncRNA	+	antisense	[506]
ABCB1	ENSG00000085563	exon/intron boundary of protein-coding gene	+	antisense	[506]
SLC22A18	ENSG00000254827	intron of protein-coding gene	+	antisense	[476]
GAB2	ENSG00000254420	antisense lncRNA	+	antisense	[506]
C18orf10	ENSG00000150477	exon of protein-coding gene	−	antisense	[476]
RUNX1	ENSG00000159216	intron of protein-coding gene	+	antisense	[506]
MCM3AP	ENSG00000215424	antisense lncRNA	+	antisense	[506]

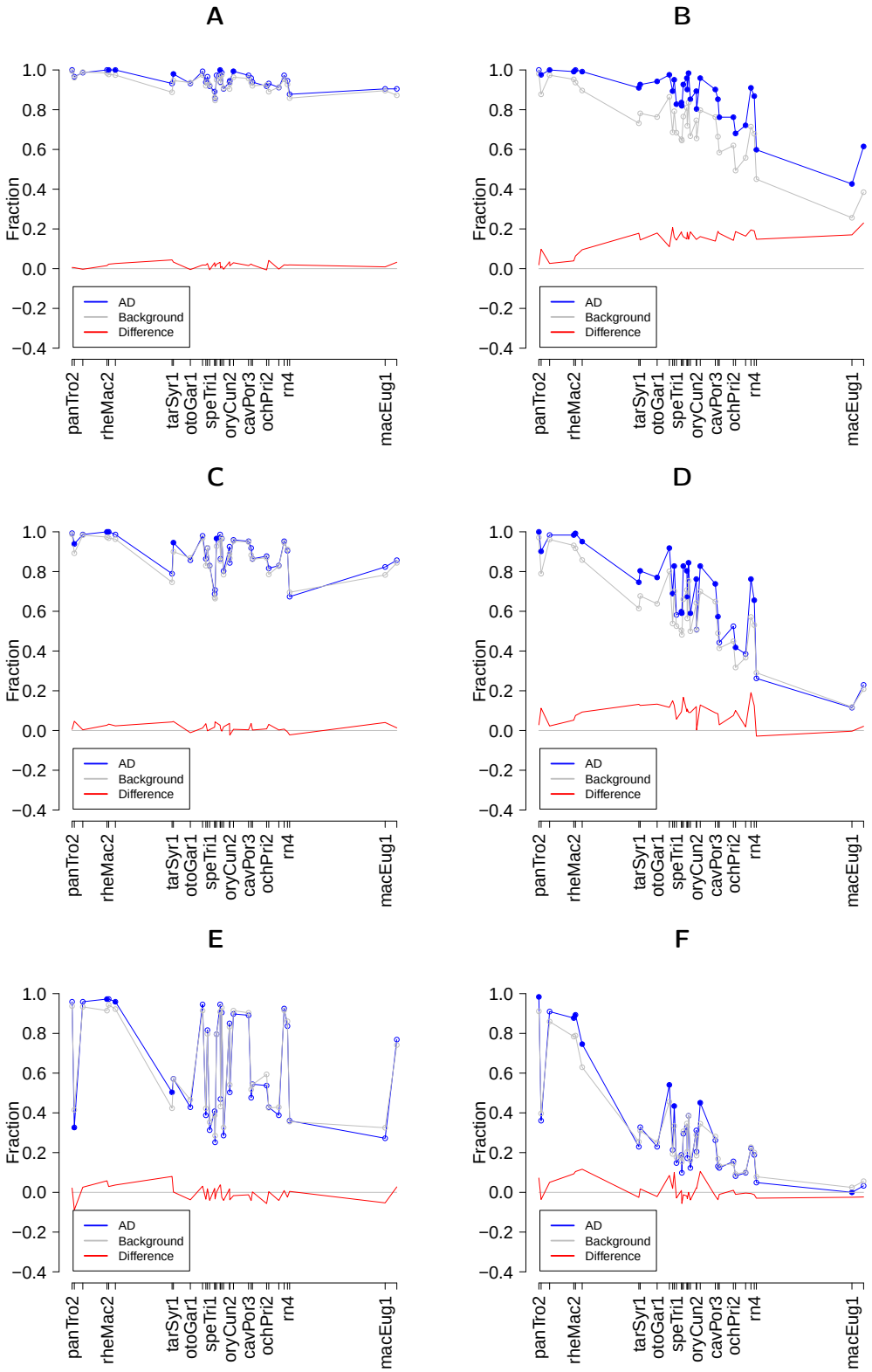


Figure C.4: *Results of the splice site conservation analyses based on the fraction of alignable genes. The figure is identical to Figure 6.6, except that the fraction of alignable genes is shown rather than the fraction of conserved genes. For more details, see Figure 6.6.*

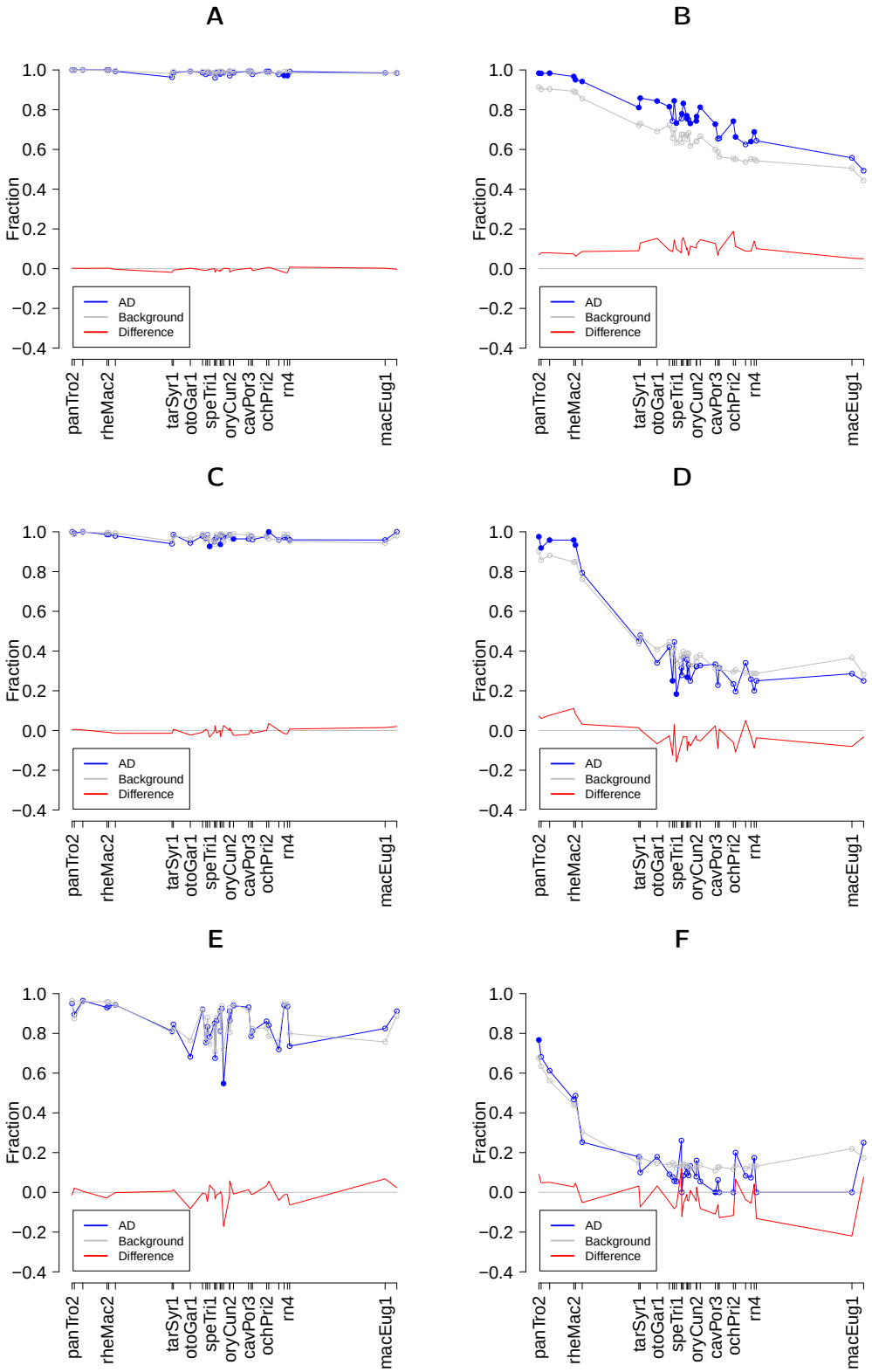


Figure C.5: *Results of the splice site conservation analyses based on the fraction of conserved genes among alignable genes. The figure is identical to Figure 6.6, except that the fraction of conserved genes among alignable genes is shown rather than the fraction of conserved genes. For more details, see Figure 6.6.*

Bibliography

[1] V. E. Russo, R. A. Martienssen, and A. D. Riggs, eds. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

[2] A. Bird. "Perceptions of epigenetics". In: *Nature* 447.7143 (2007), pp. 396–398.

[3] P. J. Denning. "Computing is a natural science". In: *Communications of the ACM* 50.7 (2007), pp. 13–18.

[4] G. Rozenberg, T. Bäck, and J. N. Kok, eds. *Handbook of Natural Computing*. Springer, 2012.

[5] L. Kari and G. Rozenberg. "The many facets of natural computing". In: *Communications of the ACM* 51.10 (October 2008), pp. 72–83.

[6] J. Kim and J. Eberwine. "RNA: state memory and mediator of cellular phenotype". In: *Trends in Cell Biology* 20.6 (2010), pp. 311–318.

[7] S. Istrail, S. B.-T. De-Leon, and E. H. Davidson. "The regulatory genome and the computer". In: *Developmental Biology* 310.2 (2007), pp. 187–195.

[8] C. Rodríguez-Caso, B. Corominas-Murtra, and R. V. Solé. "On the basic computational structure of gene regulatory networks". In: *Molecular BioSystems* 5.12 (2009), pp. 1617–1629.

[9] G. H. Moe-Behrens. "The biological microprocessor, or how to build a computer with biological parts". In: *Computational and Structural Biotechnology Journal* 7 (2013), e201304003.

[10] S. Navlakha and Z. Bar-Joseph. "Algorithms in nature: the convergence of systems biology and computational thinking". In: *Molecular Systems Biology* 7.1 (2011), p. 546.

[11] H. Abelson, D. Allen, D. Coore, C. Hanson, G. Homsy, et al. "Amorphous computing". In: *Communications of the ACM* 43.5 (2000), pp. 74–82.

[12] S. J. Prohaska, P. F. Stadler, and D. C. Krakauer. "Innovation in Gene Regulation: The Case of Chromatin Computation". In: *Journal of Theoretical Biology* 265 (2010), pp. 27–44.

[13] B. Bryant. "Chromatin Computation". In: *PLoS ONE* 7 (2012), e35703.

[14] N. Ramakrishnan, U. S. Bhalla, and J. J. Tyson. "Computing with proteins". In: *Computer* 42.1 (2009), pp. 47–56.

[15] E. T. Rolls, A. Treves, and E. T. Rolls. *Neural Networks and Brain Function*. Oxford University Press, 1998.

[16] Y. Benenson. "Biomolecular computing systems: principles, progress and potential". In: *Nature Reviews Genetics* 13.7 (2012), pp. 455–468.

[17] T. Rohlf, L. Steiner, J. Przybilla, S. Prohaska, H. Binder, et al. "Modeling the dynamic epigenome: from histone modifications towards self-organizing chromatin". In: *Epigenomics* 4.2 (2012), pp. 205–219.

[18] L. Daxinger and E. Whitelaw. "Understanding transgenerational epigenetic inheritance via the gametes in mammals". In: *Nature Reviews Genetics* 13.3 (2012), pp. 153–162.

[19] T. Kouzarides. "Chromatin modifications and their function". In: *Cell* 128.4 (2007), pp. 693–705.

[20] D. J. Owen, P. Ornaghi, J. C. Yang, N. Lowe, P. R. Evans, et al. "The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p". In: *The EMBO Journal* 19.22 (2000), pp. 6141–6149.

[21] T. Suganuma and J. L. Workman. "Crosstalk among histone modifications". In: *Cell* 135.4 (2008), pp. 604–607.

[22] M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins". In: *Nature* 410.6824 (2001), pp. 116–120.

[23] P. D. Kaufman and O. J. Rando. "Chromatin as a potential carrier of heritable information". In: *Current Opinion in Cell Biology* 22.3 (2010), pp. 284–290.

[24] R. H. Jacobson, A. G. Ladurner, D. S. King, and R. Tjian. "Structure and function of a human TAFII250 double bromodomain module". In: *Science* 288.5470 (2000), pp. 1422–1425.

[25] A. J. Bannister, P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, et al. "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain". In: *Nature* 410.6824 (2001), pp. 120–124.

[26] K. H. Hansen, A. P. Bracken, D. Pasini, N. Dietrich, S. S. Gehani, et al. "A model for transmission of the H3K27me3 epigenetic mark". In: *Nature Cell Biology* 10.11 (2008), pp. 1291–1300.

[27] M. Cockell, M. Gotta, F. Palladino, S. Martin, and S. Gasser. "Targeting Sir proteins to sites of action: a general mechanism for regulated repression". In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 63. Cold Spring Harbor Laboratory Press. 1998, pp. 401–412.

[28] S. I. Grewal and D. Moazed. "Heterochromatin and epigenetic control of gene expression". In: *Science* 301.5634 (2003), pp. 798–802.

[29] R. Margueron, N. Justin, K. Ohno, M. L. Sharpe, J. Son, et al. "Role of the Polycomb protein EED in the propagation of repressive histone marks". In: *Nature* 461.7265 (2009), pp. 762–767.

[30] D. David-Rus, S. Mukhopadhyay, J. L. Lebowitz, and A. M. Sengupta. "Inheritance of epigenetic chromatin silencing". In: *Journal of Theoretical Biology* 258.1 (2009), pp. 112–120.

[31] C. Gils, J. Wrana, and W. Salem. "A quantum spin approach to histone dynamics". In: *arXiv preprint arXiv:0912.4465* (2009).

[32] M. Sedighi and A. M. Sengupta. "Epigenetic chromatin silencing: bistability and front propagation". In: *Physical Biology* 4.4 (2007), p. 246.

[33] S. Mukhopadhyay, V. H. Nagaraj, and A. M. Sengupta. "Locus dependence in epigenetic chromatin silencing". In: *Biosystems* 102.1 (2010), pp. 49–54.

[34] M. A. Micheelsen, N. Mitarai, K. Sneppen, and I. B. Dodd. "Theory for the stability and regulation of epigenetic landscapes". In: *Physical Biology* 7.2 (2010), p. 026010.

[35] A. Dayarian and A. M. Sengupta. "Titration and hysteresis in epigenetic chromatin silencing". In: *Physical Biology* 10.3 (2013), p. 036005.

[36] I. B. Dodd, M. A. Micheelsen, K. Sneppen, and G. Thon. "Theoretical analysis of epigenetic cell memory by nucleosome modification". In: *Cell* 129 (2007), pp. 813–822.

[37] K. Sneppen, M. A. Micheelsen, and I. B. Dodd. "Ultrasensitive gene regulation by positive feedback loops in nucleosome modification". In: *Molecular Systems Biology* 4.1 (2008), p. 182.

[38] I. B. Dodd and K. Sneppen. "Barriers and silencers: A theoretical toolkit for control and containment of nucleosome-based epigenetic states". In: *Journal of Molecular Biology* 414 (2011), pp. 624–637.

[39] K. Sneppen and I. B. Dodd. "A Simple Histone Code Opens Many Paths to Epigenetics". In: *PLoS Computational Biology* 8 (2012), e1002643.

[40] E. L. Greer and Y. Shi. "Histone methylation: a dynamic mark in health, disease and inheritance". In: *Nature Reviews Genetics* 13.5 (2012), pp. 343–357.

[41] M. A. Dawson and T. Kouzarides. "Cancer epigenetics: from mechanism to therapy". In: *Cell* 150.1 (2012), pp. 12–27.

[42] J. B. Kwok. "Role of epigenetics in Alzheimer's and Parkinson's disease". In: *Biomarkers* 6.4 (2012), pp. 477–495.

[43] H. A. Irier and P. Jin. "Dynamics of DNA methylation in aging and Alzheimer's disease". In: *DNA and Cell Biology* 31.S1 (2012), S42–S48.

[44] M. Citron. "Alzheimer's disease: strategies for disease modification". In: *Nature Reviews Drug Discovery* 9.5 (2010), pp. 387–398.

[45] D. Campion, C. Dumanchin, D. Hannequin, B. Dubois, S. Belliard, et al. "Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum". In: *American Journal of Human Genetics* 65.3 (1999), p. 664.

[46] A. Wimo and M. Prince. *World Alzheimer Report 2010: the global economic impact of dementia*. Alzheimer's Disease International, 2010.

[47] L. Minati, T. Edginton, M. Grazia Bruzzone, and G. Giaccone. "Reviews: current concepts in Alzheimer's disease: a multidisciplinary review". In: *American Journal of Alzheimer's Disease and Other Dementias* 24.2 (2009), pp. 95–121.

[48] N. Zawia, D. Lahiri, and F. Cardozo–Pelaez. "Epigenetics, oxidative stress and Alzheimer's Disease". In: *Free Radical Biology & Medicine* 46.9 (2009), p. 1241.

[49] N. Coppieters and M. Dragunow. "Epigenetics in Alzheimer's Disease: a Focus on DNA Modifications". In: *Current Pharmaceutical Design* 17.31 (2012), pp. 3398–3412.

[50] J. Rao, V. Keleshian, S. Klein, and S. Rapoport. "Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients". In: *Translational Psychiatry* 2.7 (2012), e132.

[51] L. Chouliaras, B. Rutten, G. Kenis, O. Peerbooms, P. Visser, et al. "Epigenetic regulation in the pathophysiology of Alzheimer's disease". In: *Progress in Neurobiology* 90.4 (2010), p. 498.

[52] S. Massone, I. Vassallo, M. Castelnuovo, G. Fiorino, E. Gatta, et al. "RNA polymerase III drives alternative splicing of the potassium channel–interacting protein contributing to brain complexity and neurodegeneration". In: *The Journal of Cell Biology* 193.5 (2011), pp. 851–866.

[53] M. Faghihi, F. Modarresi, A. Khalil, D. Wood, B. Sahagan, et al. "Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase". In: *Nature Medicine* 14.7 (2008), pp. 723–730.

[54] O. Wapinski and H. Chang. "Long noncoding RNAs and human disease". In: *Trends in Cell Biology* 21.6 (2011), pp. 354–361.

[55] E. Heuer, R. F. Rosen, A. Cintron, and L. C. Walker. "Nonhuman primate models of Alzheimer-like cerebral proteopathy". In: *Current Pharmaceutical Design* 18.8 (2012), p. 1159.

[56] A. Toledano, M. Álvarez, A. López-Rodríguez, A. Toledano-Díaz, and C. Fernández-Verdecia. "Does Alzheimer's disease exist in all primates? Alzheimer pathology in non-human primates and its pathophysiological implications (I)". In: *Neurología (English Edition)* 27.6 (2012), pp. 354–369.

[57] J. H. Malone and B. Oliver. "Microarrays, deep sequencing and the true measure of the transcriptome". In: *BMC Biology* 9.1 (2011), p. 34.

[58] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, et al. "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses". In: *Genes & Development* 25.18 (2011), pp. 1915–1927.

[59] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, et al. "Landscape of transcription in human cells". In: *Nature* 489.7414 (2012), pp. 101–108.

[60] A. Jacquier. "The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs". In: *Nature Reviews Genetics* 10.12 (2009), pp. 833–844.

[61] M. E. Dinger, P. P. Amaral, T. R. Mercer, and J. S. Mattick. "Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications". In: *Briefings in Functional Genomics & Proteomics* 8.6 (2009), pp. 407–423.

[62] R. Holliday. "Epigenetics: a historical overview". In: *Epigenetics* 1.2 (2006), pp. 76–80.

[63] C. Waddington. "The epigenotype". In: *Endeavour* 1 (1942), pp. 18–20.

[64] R. Bonasio, S. Tu, and D. Reinberg. "Molecular signals of epigenetic states". In: *Science Signaling* 330.6004 (2010), p. 612.

[65] E. J. Richards. "Inherited epigenetic variation—revisiting soft inheritance". In: *Nature Reviews Genetics* 7.5 (2006), pp. 395–401.

[66] R. Holliday. "Mechanisms for the control of gene activity during development". In: *Biological Reviews* 65.4 (1990), pp. 431–471.

[67] S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. "An operational definition of epigenetics". In: *Genes & Development* 23.7 (2009), pp. 781–783.

[68] E. Jablonka and G. Raz. "Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution". In: *The Quarterly Review of Biology* 84.2 (2009), pp. 131–176.

[69] P. Sarkies and J. Sale. "Cellular epigenetic stability and cancer". In: *Trends in Genetics* 28.3 (2012), pp. 118–127.

[70] M. Ptashne. "On the use of the word 'epigenetic'". In: *Current Biology* 17.7 (2007), R233–R236.

[71] S. Kadauke and G. A. Blobel. "Mitotic bookmarking by transcription factors". In: *Epigenetics & Chromatin* 6.1 (2013), p. 6.

[72] A. J. Gordon, D. Satory, J. A. Halliday, and C. Herman. "Heritable Change Caused by Transient Transcription Errors". In: *PLOS Genetics* 9.6 (2013), e1003595.

[73] A. D. Goldberg, C. D. Allis, and E. Bernstein. "Epigenetics: a landscape takes shape". In: *Cell* 128.4 (2007), pp. 635–638.

[74] M. Herzog and M. O. Soyer. "Distinctive features of dinoflagellate chromatin. Absence of nucleosomes in a primitive species *Prorocentrum micans E.*" In: *European Journal of Cell Biology* 23 (1981), pp. 295–302.

[75] P. Tropberger and R. Schneider. "Scratching the (lateral) surface of chromatin regulation by histone modifications". In: *Nature Structural & Molecular Biology* 20.6 (2013), pp. 657–661.

[76] Z. Peng, M. J. Mizianty, B. Xue, L. Kurgan, and V. N. Uversky. "More than just tails: intrinsic disorder in histone proteins". In: *Molecular BioSystems* 8.7 (2012), pp. 1886–1901.

[77] A. Cascone, C. Bruelle, D. Lindholm, P. Bernardi, and O. Eriksson. "Destabilization of the outer and inner mitochondrial membranes by core and linker histones". In: *PloS one* 7.4 (2012), e35357.

[78] K. Kamieniarz, A. Izzo, M. Dundr, P. Tropberger, L. Ozretić, et al. "A dual role of linker histone H1.4 Lys 34 acetylation in transcriptional activation". In: *Genes & Development* 26.8 (2012), pp. 797–802.

[79] S.-M. Yang, B. J. Kim, L. N. Toro, and A. I. Skoultchi. "H1 linker histone promotes epigenetic silencing by regulating both DNA methylation and histone H3 methylation". In: *Proceedings of the National Academy of Sciences* 110.5 (2013), pp. 1708–1713.

[80] B. van Steensel. "Chromatin: constructing the big picture". In: *The EMBO Journal* 30.10 (2011), pp. 1885–1895.

[81] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, et al. "Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in *Drosophila* Cells". In: *Cell* 143.2 (2010), pp. 212–224.

[82] S. Henikoff and A. Shilatifard. "Histone modification: cause or cog?" In: *Trends in Genetics* 27 (2011), pp. 389–396.

[83] A. Dutta and J. L. Workman. "Nucleosome Positioning: Multiple Mechanisms toward a Unifying Goal". In: *Molecular Cell* 48.1 (2012), pp. 1–2.

[84] K. Struhl and E. Segal. "Determinants of nucleosome positioning". In: *Nature Structural & Molecular Biology* 20.3 (2013), pp. 267–273.

[85] S. Henikoff. "Nucleosome destabilization in the epigenetic regulation of gene expression". In: *Nature Reviews Genetics* 9.1 (2008), pp. 15–26.

[86] H. Kimura and P. R. Cook. "Kinetics of core histones in living human cells little exchange of H3 and H4 and some rapid exchange of H2B". In: *The Journal of Cell Biology* 153.7 (2001), pp. 1341–1354.

[87] B. M. Zee, R. S. Levin, P. A. DiMaggio, and B. A. Garcia. "Global Turnover of histone post-translational modifications and variants in human cells". In: *Epigenetics & Chromatin* 3.1 (2010), pp. 1–11.

[88] A. V. Probst, E. Dunleavy, and G. Almouzni. "Epigenetic inheritance during the cell cycle". In: *Nature Reviews Molecular Cell Biology* 10.3 (2009), pp. 192–206.

[89] R. B. Deal, J. G. Henikoff, and S. Henikoff. "Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones". In: *Science* 328.5982 (2010), pp. 1161–1164.

[90] G. Li and D. Reinberg. "Chromatin higher-order structures and gene regulation". In: *Current Opinion in Genetics & Development* 21.2 (2011), pp. 175–186.

[91] G. Fudenberg and L. A. Mirny. "Higher-order chromatin structure: bridging physics and biology". In: *Current Opinion in Genetics & Development* 22.2 (2012), pp. 115–124.

[92] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, et al. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome". In: *Nature Genetics* 39 (2007), pp. 311–318.

[93] R. Andersson, S. Enroth, A. Rada-Iglesias, C. Wadelius, and J. Komorowski. "Nucleosomes are well positioned in exons and carry characteristic histone modifications". In: *Genome Research* 19 (2009), pp. 1732–1741.

[94] S. Schwartz, E. Meshorer, and G. Ast. "Chromatin organization marks exon-intron structure". In: *Nature Structural & Molecular Biology* 16 (2009), pp. 990–995.

[95] M. D. Young, T. A. Willson, M. J. Wakefield, E. Trounson, D. J. Hilton, et al. "ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity". In: *Nucleic Acids Research* 39 (2011), pp. 7415–7427.

[96] S. C. Tippmann, R. Ivanek, D. Gaidatzis, A. Schöler, L. Hoerner, et al. "Relative contributions of different regulatory layers to steady-state mRNA levels". In: *Molecular Systems Biology* 8 (2012), p. 593.

[97] B. M. Turner. "The adjustable nucleosome: an epigenetic signaling module". In: *Trends in Genetics* 28.9 (2012), pp. 436–444.

[98] K. T. Smith and J. L. Workman. "Chromatin proteins: key responders to stress". In: *PLoS Biology* 10 (2012), e1001371.

[99] T. E. P. Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489 (2012), pp. 57–74.

[100] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, et al. "An expansive human regulatory lexicon encoded in transcription factor footprints". In: *Nature* 489.7414 (2012), pp. 83–90.

[101] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, et al. "Architecture of the human regulatory network derived from ENCODE data". In: *Nature* 489.7414 (2012), pp. 91–100.

[102] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. "The long-range interaction landscape of gene promoters". In: *Nature* 489.7414 (2012), pp. 109–113.

[103] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (2012), pp. 75–82.

[104] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, et al. "On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE". In: *Genome Biology and Evolution* 5.3 (2013), pp. 578–590.

[105] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, et al. "What is a gene, post-ENCODE? History and updated definition". In: *Genome Research* 17.6 (2007), pp. 669–681.

[106] Q. Pan, O. Shai, L. Lee, B. Frey, and B. Blencowe. "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". In: *Nature Genetics* 40.12 (2008), pp. 1413–1415.

[107] G. W. Beadle and E. L. Tatum. "Genetic control of biochemical reactions in Neurospora". In: *Proceedings of the National Academy of Sciences of the United States of America* 27.11 (1941), p. 499.

[108] A. Kalsotra and T. A. Cooper. "Functional consequences of developmentally regulated alternative splicing". In: *Nature Reviews Genetics* 12.10 (2011), pp. 715–729.

[109] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, et al. "Alternative splicing: a pivotal step between eukaryotic transcription and translation". In: *Nature Reviews Molecular Cell Biology* 14 (2013), pp. 153–165.

[110] Y. Barash, J. Calarco, W. Gao, Q. Pan, X. Wang, et al. "Deciphering the splicing code". In: *Nature* 465.7294 (2010), pp. 53–59.

[111] R. Skotheim and M. Nees. "Alternative splicing in cancer: noise, functional, or systematic?" In: *The International Journal of Biochemistry & Cell Biology* 39.7 (2007), pp. 1432–1449.

[112] C. He, F. Zhou, Z. Zuo, H. Cheng, and R. Zhou. "A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis". In: *PLOS ONE* 4.3 (2009), e4732.

[113] E. M. Rockenstein, L. McConlogue, H. Tan, M. Power, E. Masliah, et al. "Levels and alternative splicing of amyloid β protein precursor (APP) transcripts in brains of APP transgenic mice and humans with Alzheimer's disease". In: *Journal of Biological Chemistry* 270.47 (1995), pp. 28257–28267.

[114] J. R. Tollervey, Z. Wang, T. Hortobágyi, J. T. Witten, K. Zarnack, et al. "Analysis of alternative splicing associated with aging and neurodegeneration in the human brain". In: *Genome Research* 21.10 (2011), pp. 1572–1582.

[115] A. Nitsche, D. Rose, M. Fasold, and P. Stadler. "Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved". In: *in preparation* (2013).

[116] M. Tan, H. Luo, S. Lee, F. Jin, J. Yang, et al. "Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification". In: *Cell* 146.6 (2011), pp. 1016–1028.

[117] Z. Xie, J. Dai, L. Dai, M. Tan, Z. Cheng, et al. "Lysine succinylation and lysine malonylation in histones". In: *Molecular & Cellular Proteomics* 11.5 (2012), pp. 100–107.

[118] Y. Chen, R. Sprung, Y. Tang, H. Ball, B. Sangras, et al. "Lysine propionylation and butyrylation are novel post-translational modifications in histones". In: *Molecular & Cellular Proteomics* 6.5 (2007), pp. 812–819.

[119] M. Unoki, A. Masuda, N. Dohmae, K. Arita, M. Yoshimatsu, et al. "Lysyl 5-Hydroxylation, a Novel Histone Modification, by Jumonji Domain Containing 6 (JMJD6)". In: *Journal of Biological Chemistry* 288.9 (2013), pp. 6053–6062.

[120] S. Messner, M. Altmeyer, H. Zhao, A. Pozivil, B. Roschitzki, et al. "PARP1 ADP-ribosylates lysine residues of the core histone tails". In: *Nucleic Acids Research* 38.19 (2010), pp. 6350–6362.

[121] K. Sakabe, Z. Wang, and G. W. Hart. "β-N-acetylglucosamine (O-GlcNAc) is part of the histone code". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.46 (2010), pp. 19915–19920.

[122] B. M. Turner. "Reading signals on the nucleosome with a new nomenclature for modified histones". In: *Nature Structural & Molecular Biology* 12.2 (2005), pp. 110–112.

[123] B. György, E. Tóth, E. Tarcsa, A. Falus, and E. I. Buzás. "Citrullination: a posttranslational modification in health and disease". In: *The International Journal of Biochemistry & Cell Biology* 38.10 (2006), pp. 1662–1677.

[124] J. Park, Y. Chen, D. X. Tishkoff, C. Peng, M. Tan, et al. "SIRT5-Mediated Lysine Desuccinylation Impacts Diverse Metabolic Pathways". In: *Molecular Cell* 50.6 (2013), pp. 919–930.

[125] B. C. Smith, W. C. Hallows, and J. M. Denu. "Mechanisms and molecular probes of sirtuins". In: *Chemistry & Biology* 15.10 (2008), pp. 1002–1013.

[126] P. Voigt, G. LeRoy, W. Drury III, B. Zee, J. Son, et al. "Asymmetrically Modified Nucleosomes". In: *Cell* 151.1 (2012), pp. 181–193.

[127] V. Tran, C. Lim, J. Xie, and X. Chen. "Asymmetric Division of *Drosophila* Male Germline Stem Cell Shows Asymmetric Histone Distribution". In: *Science* 338.6107 (2012), pp. 679–682.

[128] T. Banerjee and D. Chakravarti. "A peek into the complex realm of histone phosphorylation". In: *Molecular and Cellular Biology* 31.24 (2011), pp. 4858–4873.

[129] J. Füllgrabe, E. Kavanagh, and B. Joseph. "Histone onco-modifications". In: *Oncogene* 30.31 (2011), pp. 3391–3403.

[130] B. Li, M. Carey, and J. L. Workman. "The role of chromatin during transcription". In: *Cell* 128.4 (2007), pp. 707–719.

[131] K. M. Miller and S. P. Jackson. "Histone marks: repairing DNA breaks within the context of chromatin". In: *Biochemical Society Transactions* 40 (2012), pp. 370–376.

[132] E. J. Wagner and P. B. Carpenter. "Understanding the language of Lys36 methylation at histone H3". In: *Nature Reviews Molecular Cell Biology* 13.2 (2012), pp. 115–126.

[133] R. Margueron, P. Trojer, and D. Reinberg. "The key to development: interpreting the histone code?" In: *Current Opinion in Genetics & Development* 15.2 (2005), pp. 163–176.

[134] Y. H. Woo and W.-H. Li. "Evolutionary conservation of histone modifications in mammals". In: *Molecular Biology and Evolution* 29.7 (2012), pp. 1757–1767.

[135] M. S. Cosgrove. "Histone proteomics and the epigenetic regulation of nucleosome mobility". In: *Expert Review of Proteomics* 4.4 (2007), pp. 465–478.

[136] A. R. Pengelly, Ö. Copur, H. Jäckle, A. Herzig, and J. Müller. "A Histone Mutant Reproduces the Phenotype Caused by Loss of Histone-Modifying Factor Polycomb". In: *Science* 339.6120 (2013), pp. 698–699.

[137] M. Hödl and K. Basler. "Transcription in the Absence of Histone H3.2 and H3K4 Methylation". In: *Current Biology* 23.4 (2012), pp. 2253–2257.

[138] Z. Ge, D. Nair, X. Guan, N. Rastogi, M. A. Freitas, et al. "Sites of acetylation on newly synthesized histone H4 are required for chromatin assembly and DNA damage response signaling". In: *Molecular and Cellular Biology* 33.16 (2013), pp. 3286–3298.

[139] C. Huang, M. Xu, and B. Zhu. "Epigenetic inheritance mediated by histone lysine methylation: maintaining transcriptional states without the precise restoration of marks?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1609 (2013), p. 20110332.

[140] D. Moazed. "Mechanisms for the Inheritance of Chromatin States". In: *Cell* 146 (2011), pp. 510–518.

[141] A. Bannister and T. Kouzarides. "Regulation of chromatin by histone modifications". In: *Cell Research* 21.3 (2011), pp. 381–395.

[142] T. Mikkelsen, M. Ku, D. Jaffe, B. Issac, E. Lieberman, et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153 (2007), pp. 553–560.

[143] B. D. Strahl and C. D. Allis. "The language of covalent histone modifications". In: *Nature* 403.6765 (2000), pp. 41–45.

[144] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, et al. "Combinatorial patterns of histone acetylations and methylations in the human genome". In: *Nature Genetics* 40.7 (2008), pp. 897–903.

[145] Z. Wang and D. J. Patel. "Combinatorial readout of dual histone modifications by paired chromatin-associated modules". In: *The Journal of Biological Chemistry* 286 (2011), pp. 18363–18368.

[146] A. Lindroth, D. Shultis, Z. Jasencakova, J. Fuchs, L. Johnson, et al. "Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3". In: *The EMBO Journal* 23.21 (2004), pp. 4146–4155.

[147] S. M. Fuchs, K. Krajewski, R. W. Baker, V. L. Miller, and B. D. Strahl. "Influence of combinatorial histone modifications on antibody and effector protein recognition". In: *Current Biology* 21.1 (2011), pp. 53–58.

[148] A. J. Ruthenburg, H. Li, T. A. Milne, S. Dewell, R. K. McGinty, et al. "Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions". In: *Cell* 145.5 (2011), pp. 692–706.

[149] C. Musselman, M. Lalonde, J. Côté, and T. Kutateladze. "Perceiving the epigenetic landscape through histone readers". In: *Nature Structural & Molecular Biology* 19.12 (2012), pp. 1218–1227.

[150] D. Molina-Serrano, V. Schiza, and A. Kirmizis. "Cross-talk among epigenetic modifications: lessons from histone arginine methylation". In: *Biochemical Society Transactions* 41.3 (2013), pp. 751–759.

[151] L. Wu, S. Y. Lee, B. Zhou, U. T. Nguyen, T. W. Muir, et al. "ASH2L Regulates Ubiquitylation Signaling to MLL: *trans*-Regulation of H3 K4 Methylation in Higher Eukaryotes". In: *Molecular Cell* 49.6 (2013), pp. 1108–1120.

[152] J. Kim, J. Kim, R. K. McGinty, U. T. Nguyen, T. W. Muir, et al. "The n-SET Domain of Set1 Regulates H2B Ubiquitylation-Dependent H3K4 Methylation". In: *Molecular Cell* 49.6 (2013), pp. 1121–33.

[153] A. Zippo, R. Serafini, M. Rocchigiani, S. Pennacchini, A. Krepelova, et al. "Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation". In: *Cell* 138.6 (2009), pp. 1122–1136.

[154] W. Fischle, B. S. Tseng, H. L. Dormann, B. M. Ueberheide, B. A. Garcia, et al. "Regulation of HP1–chromatin binding by histone H3 methylation and phosphorylation". In: *Nature* 438.7071 (2005), pp. 1116–1122.

[155] B. Mateescu, B. Bourachot, C. Rachez, V. Ogryzko, and C. Muchardt. "Regulation of an inducible promoter by an HP1β–HP1γ switch". In: *EMBO Reports* 9.3 (2008), pp. 267–272.

[156] P. Cheung, K. G. Tanner, W. L. Cheung, P. Sassone-Corsi, J. M. Denu, et al. "Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation". In: *Molecular Cell* 5.6 (2000), pp. 905–915.

[157] W.-S. Lo, R. C. Trievel, J. R. Rojas, L. Duggan, J.-Y. Hsu, et al. "Phosphorylation of serine 10 in histone H3 is functionally linked *in vitro* and *in vivo* to Gcn5-mediated acetylation at lysine 14". In: *Molecular Cell* 5.6 (2000), pp. 917–926.

[158] J. A. Latham, R. J. Chosed, S. Wang, and S. Y. Dent. "Chromatin signaling to kinetochores: transregulation of Dam1 methylation by histone H2B ubiquitination". In: *Cell* 146.5 (2011), pp. 709–719.

[159] P. N. I. Lau and P. Cheung. "Elucidating combinatorial histone modifications and crosstalks by coupling histone-modifying enzyme with biotin ligase activity". In: *Nucleic Acids Research* 41.3 (2013), e49.

[160] X. Guan, N. Rastogi, M. R. Parthun, and M. A. Freitas. "Discovery of Histone Modification Crosstalk Networks by SILAC Mass Spectrometry". In: *Molecular & Cellular Proteomics* 12.8 (2013), pp. 2048–2059.

[161] B. M. Turner. "Environmental sensing by chromatin: An epigenetic contribution to evolutionary change". In: *FEBS Letters* 585.13 (2011), pp. 2032–2040.

[162] R. Scully. "A histone code for DNA repair". In: *Nature Reviews Molecular Cell Biology* 11.3 (2010), pp. 164–164.

[163] T. Suganuma and J. Workman. "Signals and combinatorial functions of histone modifications". In: *Annual Review of Biochemistry* 80 (2011), pp. 473–499.

[164] G. E. Zentner and S. Henikoff. "Regulation of nucleosome dynamics by histone modifications". In: *Nature Structural & Molecular Biology* 20.3 (2013), pp. 259–266.

[165] R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, et al. "Regulation of alternative splicing by histone modifications". In: *Science* 327.5968 (2010), pp. 996–1000.

[166] F. Li, G. Mao, D. Tong, J. Huang, L. Gu, et al. "The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through Its Interaction with MutSα". In: *Cell* 153.3 (2013), pp. 590–600.

[167] H. Pei, L. Zhang, K. Luo, Y. Qin, M. Chesi, et al. "MMSET regulates histone H4K20 methylation and 53BP1 accumulation at DNA damage sites". In: *Nature* 470.7332 (2011), pp. 124–128.

[168] S. Jørgensen, G. Schotta, and C. S. Sørensen. "Histone H4 Lysine 20 methylation: key player in epigenetic regulation of genomic integrity". In: *Nucleic Acids Research* 41.5 (2013), pp. 2797–2806.

[169] J. M. Schulze, J. Jackson, S. Nakanishi, J. M. Gardner, T. Hentrich, et al. "Linking cell cycle to histone modifications: SBF and H2B monoubiquitination machinery and cell-cycle regulation of H3K79 dimethylation". In: *Molecular Cell* 35.5 (2009), pp. 626–641.

[170] D. De Vos, F. Frederiks, M. Terweij, T. van Welsem, K. F. Verzijlbergen, et al. "Progressive methylation of ageing histones by Dot1 functions as a timer". In: *EMBO Reports* 12.9 (2011), pp. 956–962.

[171] F. Wang and J. Higgins. "Histone modifications and mitosis: countermarks, landmarks, and bookmarks". In: *Trends in Cell Biology* 23.4 (2012), pp. 175–184.

[172] M. Xu, C. Long, X. Chen, C. Huang, S. Chen, et al. "Partitioning of Histone H3-H4 Tetramers During DNA Replication–Dependent Chromatin Assembly". In: *Science* 328.5974 (2010), pp. 94–98.

[173] T. K. Barth and A. Imhof. "Fast signals and slow marks: the dynamics of histone modifications". In: *Trends in Biochemical Sciences* 35 (2010), pp. 618–626.

[174] N. J. Francis. "Gene Regulation: Implications of Histone Dispersal Patterns for Epigenetics". In: *Current Biology* 21.17 (2011), R659–R661.

[175] H. Santos-Rosa, A. Kirmizis, C. Nelson, T. Bartke, N. Saksouk, et al. "Histone H3 tail clipping regulates gene expression". In: *Nature Structural & Molecular Biology* 16.1 (2008), pp. 17–22.

[176] E. M. Duncan, T. L. Muratore-Schroeder, R. G. Cook, B. A. Garcia, J. Shabanowitz, et al. "Cathepsin L proteolytically processes histone H3 during mouse embryonic stem cell differentiation". In: *Cell* 135.2 (2008), pp. 284–294.

[177] P. Mandal, N. Verma, S. Chauhan, and R. S. Tomar. "Unexpected histone H3 tail clipping activity of Glutamate dehydrogenase". In: *Journal of Biological Chemistry* 288.26 (2013), pp. 18743–18757.

[178] P. Sarkies and J. E. Sale. "Propagation of histone marks and epigenetic memory during normal and interrupted DNA replication". In: *Cellular and Molecular Life Sciences* (2011), pp. 1–20.

[179] F. Golebiowski, I. Matic, M. H. Tatham, C. Cole, Y. Yin, et al. "System-wide changes to SUMO modifications in response to heat shock". In: *Science Signaling* 2.72 (2009), ra24.

[180] O. J. Rando. "Combinatorial complexity in chromatin structure and function: revisiting the histone code". In: *Current Opinion in Genetics & Development* 22.2 (2012), pp. 148–155.

[181] B. Zhu and D. Reinberg. "Epigenetic inheritance: uncontested?" In: *Cell Research* 21.3 (2011), pp. 435–441.

[182] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398 (2012), pp. 376–380.

[183] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398 (2012), pp. 381–385.

[184] E. de Wit, B. A. Bouwman, Y. Zhu, P. Klous, E. Splinter, et al. "The pluripotent genome in three dimensions is shaped around pluripotency factors". In: *Nature* 501 (2013), pp. 227–231.

[185] R. Margueron and D. Reinberg. "Chromatin structure and the inheritance of epigenetic information". In: *Nature Reviews Genetics* 11.4 (2010), pp. 285–296.

[186] C. Maison, D. Bailly, A. H. Peters, J.-P. Quivy, D. Roche, et al. "Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component". In: *Nature Genetics* 30.3 (2002), pp. 329–334.

[187] A. Tedeschi, G. Wutz, S. Huet, M. Jaritz, A. Wuensche, et al. "Wapl is an essential regulator of chromatin structure and chromosome segregation". In: *Nature* 501.7468 (2013), pp. 564–568.

[188] T. Mondal, M. Rasmussen, G. Pandey, A. Isaksson, and C. Kanduri. "Characterization of the RNA content of chromatin". In: *Genome Research* 20.7 (2010), pp. 899–907.

[189] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, et al. "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals". In: *Nature* 458.7235 (March 2009), pp. 223–227.

[190] M. Guttman and J. L. Rinn. "Modular regulatory principles of large non-coding RNAs". In: *Nature* 482.7385 (2012), pp. 339–346.

[191] L. Nie, H.-J. Wu, J.-M. Hsu, S.-S. Chang, A. M. LaBaff, et al. "Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer". In: *American Journal of Translational Research* 4.2 (2012), p. 127.

[192] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, et al. "Circular RNAs are a large class of animal RNAs with regulatory potency". In: *Nature* 495.7441 (2013), pp. 333–338.

[193] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, et al. "Natural RNA circles function as efficient microRNA sponges". In: *Nature* 495.7441 (2013), pp. 384–388.

[194] T. R. Mercer and J. S. Mattick. "Structure and function of long noncoding RNAs in epigenetic regulation". In: *Nature Structural & Molecular Biology* 20.3 (2013), pp. 300–307.

[195] K. Van Roosbroeck, J. Pollet, and G. A. Calin. "miRNAs and long noncoding RNAs as biomarkers in human diseases". In: *Expert Review of Molecular Diagnostics* 13.2 (2013), pp. 183–204.

[196] I. A. Vergara, N. Erho, T. J. Triche, M. Ghadessi, A. Crisan, et al. "Genomic "Dark Matter" in prostate cancer: Exploring the clinical utility of ncRNA as biomarkers". In: *Frontiers in Genetics* 3 (2012), p. 23.

[197] C. A. Edwards and A. C. Ferguson-Smith. "Mechanisms regulating imprinted genes in clusters". In: *Current Opinion in Cell Biology* 19.3 (2007), pp. 281–289.

[198] A. Wutz. "Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation". In: *Nature Reviews Genetics* 12.8 (2011), pp. 542–553.

[199] J. Lee. "Epigenetic Regulation by Long Noncoding RNAs". In: *Science* 338.6113 (2012), pp. 1435–1439.

[200] M. Magistri, M. A. Faghihi, G. St Laurent III, and C. Wahlestedt. "Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts". In: *Trends in Genetics* 28.8 (2012), pp. 389–396.

[201] A. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, et al. "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.28 (2009), pp. 11667–11672.

[202] J. Zhao, B. K. Sun, J. A. Erwin, J.-J. Song, and J. T. Lee. "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome". In: *Science* 322.5902 (2008), pp. 750–756.

[203] K. L. Yap, S. Li, A. M. Muñoz-Cabello, S. Raguz, L. Zeng, et al. "Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a". In: *Molecular Cell* 38.5 (2010), pp. 662–674.

[204] M. E. Dinger, P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, et al. "Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation". In: *Genome Research* 18.9 (2008), pp. 1433–1445.

[205] M. J. Koziol and J. L. Rinn. "RNA traffic control of chromatin complexes". In: *Current Opinion in Genetics & Development* 20.2 (2010), pp. 142–148.

[206] F. Lai, U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, et al. "Activating RNAs associate with Mediator to enhance chromatin architecture and transcription". In: *Nature* 494.7438 (2013), pp. 497–501.

[207] M. T. Knuesel, K. D. Meyer, A. J. Donner, J. M. Espinosa, and D. J. Taatjes. "The human CDK8 subcomplex is a histone kinase that requires Med12 for activity and can function independently of mediator". In: *Molecular and Cellular Biology* 29.3 (2009), pp. 650–661.

[208] C. Keller, R. Kulasegaran-Shylini, Y. Shimada, H.-R. Hotz, and M. Bühler. "Noncoding RNAs prevent spreading of a repressive histone mark". In: *Nature Structural & Molecular Biology* 20 (2013), pp. 994–1000.

[209] B. A. Buckley, K. B. Burkhart, S. G. Gu, G. Spracklin, A. Kershner, et al. "A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality". In: *Nature* 489.7416 (2012), pp. 447–451.

[210] T. A. Volpe, C. Kidner, I. M. Hall, G. Teng, S. I. Grewal, et al. "Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi". In: *Science* 297.5588 (2002), pp. 1833–1837.

[211] R. C. Allshire and G. H. Karpen. "Epigenetic regulation of centromeric chromatin: old dogs, new tricks?" In: *Nature Reviews Genetics* 9.12 (2008), pp. 923–937.

[212] C. D. Malone and G. J. Hannon. "Small RNAs as guardians of the genome". In: *Cell* 136.4 (2009), pp. 656–668.

[213] M. Gonzalez and F. Li. "DNA replication, RNAi and epigenetic inheritance". In: *Epigenetics* 7.1 (2012), pp. 14–19.

[214] S. G. Gu, J. Pak, S. Guang, J. M. Maniar, S. Kennedy, et al. "Amplification of siRNA in *Caenorhabditis elegans* generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint". In: *Nature Genetics* 44.2 (2012), pp. 157–164.

[215] F. Cernilogar, M. Onorati, G. Kothe, A. Burroughs, K. Parsi, et al. "Chromatin-associated RNA interference components contribute to transcriptional regulation in *Drosophila*". In: *Nature* 480.7377 (2011), pp. 391–395.

[216] B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli. "Genome Regulation by Polycomb and Trithorax Proteins". In: *Cell* 128.4 (2007), pp. 735–745.

[217] Y.-i. Tsukada, J. Fang, H. Erdjument-Bromage, M. E. Warren, C. H. Borchers, et al. "Histone demethylation by a family of JmjC domain-containing proteins". In: *Nature* 439.7078 (2005), pp. 811–816.

[218] M. Yun, J. Wu, J. Workman, and B. Li. "Readers of histone modifications". In: *Cell Research* 21.4 (2011), pp. 564–578.

[219] K. E. Gardner, C. D. Allis, and B. D. Strahl. "Operating on chromatin, a colorful language where context matters". In: *Journal of Molecular Biology* 409 (2011), pp. 36–46.

[220] F. Frederiks, M. Tzouros, G. Oudgenoeg, T. van Welsem, M. Fornerod, et al. "Nonprocessive methylation by Dot1 leads to functional redundancy of histone H3K79 methylation states". In: *Nature Structural & Molecular Biology* 15.6 (2008), pp. 550–557.

[221] S. C. Wu and Y. Zhang. "Active DNA demethylation: many roads lead to Rome". In: *Nature Reviews Molecular Cell Biology* 11.9 (2010), pp. 607–620.

[222] A. N. Scharf and A. Imhof. "Every methyl counts–epigenetic calculus". In: *FEBS Letters* 585.13 (2011), pp. 2001–2007.

[223] A. J. Ruthenburg, H. Li, D. J. Patel, and C. D. Allis. "Multivalent engagement of chromatin modifications by linked binding modules". In: *Nature Reviews Molecular Cell Biology* 8.12 (2007), pp. 983–994.

[224] J. Morinière, S. Rousseaux, U. Steuerwald, M. Soler-López, S. Curtet, et al. "Cooperative binding of two acetylation marks on a histone tail by a single bromodomain". In: *Nature* 461.7264 (2009), pp. 664–668.

[225] A. Fradet-Turcotte, M. D. Canny, C. Escribano-Díaz, A. Orthwein, C. C. Leung, et al. "53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark". In: *Nature* 499 (2013), pp. 50–54.

[226] S. Pu, A. L. Turinsky, J. Vlasblom, T. On, X. Xiong, et al. "Expanding the landscape of chromatin modification (CM)-related functional domains and genes in human". In: *PLOS ONE* 5.11 (2010), e14122.

[227] E. G. Clements, H. P. Mohammad, B. R. Leadem, H. Easwaran, Y. Cai, et al. "DNMT1 modulates gene expression without its catalytic activity partially through its interactions with histone-modifying enzymes". In: *Nucleic Acids Research* 40.10 (2012), pp. 4334–4346.

[228] K. Mahajan and N. P. Mahajan. "WEE1 tyrosine kinase, a novel epigenetic modifier". In: *Trends in Genetics* 29.7 (2013), pp. 394–402.

[229] B. M. Turner. "Epigenetic responses to environmental change and their evolutionary implications". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1534 (2009), pp. 3403–3418.

[230] S. Kuroki, S. Matoba, M. Akiyoshi, Y. Matsumura, H. Miyachi, et al. "Epigenetic regulation of mouse sex determination by the histone demethylase Jmjd1a". In: *Science* 341.6150 (2013), pp. 1106–1109.

[231] M. Rolando, S. Sanulli, C. Rusniok, L. Gomez-Valero, C. Bertholet, et al. "*Legionella pneumophila* Effector RomA Uniquely Modifies Host Chromatin to Repress Gene Expression and Promote Intracellular Bacterial Replication". In: *Cell Host & Microbe* 13.4 (2013), pp. 395–405.

[232] S. Zaidi, M. Choi, H. Wakimoto, L. Ma, J. Jiang, et al. "*De novo* mutations in histone-modifying genes in congenital heart disease". In: *Nature* 498.7453 (2013), pp. 220–223.

[233] B. Mar, L. Bullinger, E. Basu, K. Schlis, L. Silverman, et al. "Sequencing histone-modifying enzymes identifies UTX mutations in acute lymphoblastic leukemia". In: *Leukemia* 26.8 (2012), pp. 1881–1883.

[234] D. Fachinetti, H. D. Folco, Y. Nechemia-Arbely, L. P. Valente, K. Nguyen, et al. "A two-step mechanism for epigenetic specification of centromere identity and function". In: *Nature Cell Biology* 15.9 (2013), pp. 1056–66.

[235] P. Talbert and S. Henikoff. "Histone variants – ancient wrap artists of the epigenome". In: *Nature Reviews Molecular Cell Biology* 11.4 (2010), pp. 264–275.

[236] G. Yuan and B. Zhu. "Histone variants and epigenetic inheritance". In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819.3 (2012), pp. 222–229.

[237] B. E. Black and D. W. Cleveland. "Epigenetic centromere propagation and the nature of CENP-A nucleosomes". In: *Cell* 144.4 (2011), pp. 471–479.

[238] S. Henikoff and T. Furuyama. "Epigenetic inheritance of centromeres". In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 75. Cold Spring Harbor Laboratory Press. 2010, pp. 51–60.

[239] S. B. Hake and C. D. Allis. "Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis"". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.17 (2006), pp. 6428–6435.

[240] N. Bhutani, D. M. Burns, and H. M. Blau. "DNA demethylation dynamics". In: *Cell* 146.6 (2011), pp. 866–872.

[241] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, et al. "Human DNA methylomes at base resolution show widespread epigenomic differences". In: *Nature* 462.7271 (2009), pp. 315–322.

[242] P. A. Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492.

[243] A. Nishiyama, L. Yamaguchi, J. Sharif, Y. Johmura, T. Kawamura, et al. "Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication". In: *Nature* 502 (2013), pp. 249–253.

[244] H. Wu and Y. Zhang. "Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells". In: *Cell Cycle* 10.15 (2011), pp. 2428–2436.

[245] J. A. Hackett, R. Sengupta, J. J. Zylicz, K. Murakami, C. Lee, et al. "Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine". In: *Science* 339.6118 (2013), pp. 448–452.

[246] M. R. Branco, G. Ficz, and W. Reik. "Uncovering the role of 5-hydroxymethylcytosine in the epigenome". In: *Nature Reviews Genetics* 13.1 (2011), pp. 7–13.

[247] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, et al. "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine". In: *Science* 333.6047 (2011), pp. 1300–1303.

[248] Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, et al. "Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA". In: *Science* 333.6047 (2011), pp. 1303–1307.

[249] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, et al. "The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA". In: *Angewandte Chemie* 123.31 (2011), pp. 7146–7150.

[250] M. W. Kellinger, C.-X. Song, J. Chong, X.-Y. Lu, C. He, et al. "5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription". In: *Nature Structural & Molecular Biology* 19.8 (2012), pp. 831–833.

[251] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, et al. "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types". In: *Genome Research* 20.6 (2010), pp. 761–770.

[252] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, et al. "Global reorganization of replication domains during embryonic stem cell differentiation". In: *PLoS Biology* 6.10 (2008), e245.

[253] R. Hand. "Eucaryotic DNA: organization of the genome for replication". In: *Cell* 15.2 (1978), p. 317.

[254] C. Alabert and A. Groth. "Chromatin replication and epigenome maintenance". In: *Nature Reviews Molecular Cell Biology* 13.3 (2012), pp. 153–167.

[255] Y. Lorch, B. Maier-Davis, and R. D. Kornberg. "Chromatin remodeling by nucleosome disassembly *in vitro*". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.9 (2006), pp. 3090–3093.

[256] C. Huang, Z. Zhang, M. Xu, Y. Li, Z. Li, et al. "H3.3-H4 Tetramer Splitting Events Feature Cell-Type Specific Enhancers". In: *PLOS Genetics* 9.6 (2013), e1003558.

[257] V. Jackson and R. Chalkley. "Histone segregation of replicating chromatin". In: *Biochemistry* 24 (1985), pp. 6930–6938.

[258] A. Corpet and G. Almouzni. "Making copies of chromatin: the challenge of nucleosomal organization and epigenetic information". In: *Trends in Cell Biology* 19.1 (2009), pp. 29–41.

[259] A. T. Annunziato. "Split decision: what happens to nucleosomes during DNA replication?" In: *Journal of Biological Chemistry* 280.13 (2005), pp. 12065–12068.

[260] M. Radman-Livaja, K. F. Verzijlbergen, A. W. Weiner, T. van Welsem, N. Friedman, et al. "Patterns and mechanisms of ancestral histone protein inheritance in budding yeast". In: *PLoS Biology* 9.6 (2011), e1001075.

[261] S. Petruk, Y. Sedkov, D. M. Johnston, J. W. Hodgson, K. L. Black, et al. "TrxG and PcG Proteins but Not Methylated Histones Remain Associated with DNA through Replication". In: *Cell* 150.5 (2012), pp. 922–933.

[262] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, et al. "Genome-scale identification of nucleosome positions in *S. cerevisiae*". In: *Science* 309.5734 (2005), pp. 626–630.

[263] M. J. Luteijn and R. F. Ketting. "PIWI-interacting RNAs: from generation to transgenerational epigenetics". In: *Nature Reviews Genetics* 14 (2013), pp. 523–534.

[264] J. Kiani, V. Grandjean, R. Liebers, F. Tuorto, H. Ghanbarian, et al. "RNA–Mediated Epigenetic Heredity Requires the Cytosine Methyltransferase Dnmt2". In: *PLOS Genetics* 9.5 (2013), e1003498.

[265] A. Molnar, C. W. Melnyk, A. Bassett, T. J. Hardcastle, R. Dunn, et al. "Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells". In: *Science* 328.5980 (2010), pp. 872–875.

[266] O. J. Rando and K. J. Verstrepen. "Timescales of genetic and epigenetic inheritance". In: *Cell* 128.4 (2007), pp. 655–668.

[267] R. Bonduriansky. "Rethinking heredity, again". In: *Trends in Ecology & Evolution* 27.6 (2012), pp. 330–336.

[268] I. Whitehouse and D. Smith. "Chromatin dynamics at the replication fork: there's more to life than histones". In: *Current Opinion in Genetics & Development* 23.2 (2013), pp. 140–146.

[269] C. Guetg, F. Scheifele, F. Rosenthal, M. O. Hottiger, and R. Santoro. "Inheritance of silent rDNA chromatin is mediated by PARP1 via noncoding RNA". In: *Molecular Cell* 45.6 (2012), pp. 790–800.

[270] J. J. Pesavento, H. Yang, N. L. Kelleher, and C. A. Mizzen. "Certain and progressive methylation of histone H4 at lysine 20 during the cell cycle". In: *Molecular and Cellular Biology* 28.1 (2008), pp. 468–486.

[271] Y. Katan-Khaykovich and K. Struhl. "Heterochromatin formation involves changes in histone modifications over multiple cell generations". In: *The EMBO Journal* 24.12 (2005), pp. 2138–2149.

[272] T. Moss, F. Langlois, T. Gagnon-Kugler, and V. Stefanovsky. "A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis". In: *Cellular and Molecular Life Sciences* 64.1 (2007), pp. 29–49.

[273] K. D. Sarge and O.-K. Park-Sarge. "Gene bookmarking: keeping the pages open". In: *Trends in Biochemical Sciences* 30.11 (2005), pp. 605–610.

[274] R. Zhao, T. Nakamura, Y. Fu, Z. Lazar, and D. L. Spector. "Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation". In: *Nature Cell Biology* 13.11 (2011), pp. 1295–1304.

[275] S. K. Zaidi, D. W. Young, M. A. Montecino, J. B. Lian, A. J. Van Wijnen, et al. "Mitotic bookmarking of genes: a novel dimension to epigenetic control". In: *Nature Reviews Genetics* 11.8 (2010), pp. 583–589.

[276] G. A. Blobel, S. Kadauke, E. Wang, A. W. Lau, J. Zuber, et al. "A reconfigured pattern of MLL occupancy within mitotic chromatin promotes rapid transcriptional reactivation following mitotic exit". In: *Molecular Cell* 36.6 (2009), pp. 970–983.

[277] N. A. Hathaway, O. Bell, C. Hodges, E. L. Miller, D. S. Neel, et al. "Dynamics and memory of heterochromatin in living cells". In: *Cell* 149.7 (2012), pp. 1447–1460.

[278] M. Grunstein. "Yeast heterochromatin: regulation of its assembly and inheritance by histones". In: *Cell* 93.3 (1998), pp. 325–328.

[279] B. M. Turner. "Histone acetylation as an epigenetic determinant of long-term transcriptional competence". In: *Cellular and Molecular Life Sciences* 54.1 (1998), pp. 21–31.

[280] A. Angel, J. Song, C. Dean, and M. Howard. "A Polycomb-based switch underlying quantitative epigenetic memory". In: *Nature* 476.7358 (2011), pp. 105–108.

[281] D. E. Koshland and K. Hamadani. "Proteomics and models for enzyme cooperativity". In: *Journal of Biological Chemistry* 277.49 (2002), pp. 46841–46844.

[282] N. Blüthgen, S. Legewie, H. Herzel, and B. Kholodenko. "Mechanisms generating ultrasensitivity, bistability, and oscillations in signal transduction". In: *Introduction to Systems Biology*. Springer, 2007, pp. 282–299.

[283] J. J. Tyson, K. C. Chen, and B. Novak. "Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell". In: *Current Opinion in Cell Biology* 15.2 (2003), pp. 221–231.

[284] R. Thomas and M. Kaufman. "Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 11.1 (2001), pp. 170–179.

[285] W. Xiong and J. E. Ferrell. "A positive-feedback-based bistable 'memory module' that governs a cell fate decision". In: *Nature* 426.6965 (2003), pp. 460–465.

[286] W. K. Smits, O. P. Kuipers, and J.-W. Veening. "Phenotypic variation in bacteria: the role of feedback regulation". In: *Nature Reviews Microbiology* 4.4 (2006), pp. 259–271.

[287] J. E. Ferrell and W. Xiong. "Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 11.1 (2001), pp. 227–236.

[288] C. Hodges and G. Crabtree. "Dynamics of inherently bounded histone modification domains". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.33 (2012), pp. 13296–13301.

[289] P. A. Steffen, J. P. Fonseca, and L. Ringrose. "Epigenetics meets mathematics: Towards a quantitative understanding of chromatin biology". In: *BioEssays* 34.10 (2012), pp. 901–913.

[290] S. G. Akl. *WHAT IS COMPUTATION?* Tech. rep. School of Computing, Queen's University, 2013.

[291] Z. W. Pylyshyn. *Computation and Cognition*. Cambridge Univ Press, 1984.

[292] D. B. Fogel. "What is evolutionary computation?" In: *Spectrum, IEEE* 37.2 (2000), pp. 26–28.

[293] B. J. Copeland. "What is computation?" In: *Synthese* 108.3 (1996), pp. 335–359.

[294] G. Piccinini and A. Scarantino. "Computation vs. information processing: why their difference matters to cognitive science". In: *Studies in History and Philosophy of Science Part A* 41.3 (2010), pp. 237–246.

[295] A. M. Turing. "On computable numbers, with an application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* 42.2 (1936), pp. 230–265.

[296] G. Piccinini and A. Scarantino. "Information processing, computation, and cognition". In: *Journal of Biological Physics* 37.1 (2011), pp. 1–38.

[297] N. Fresco. "Computation as Information Processing". In: *Physical Computation and Cognitive Science*. Springer, 2014, pp. 133–166.

[298] G. Păun. "Computing with membranes". In: *Journal of Computer and System Sciences* 61.1 (2000), pp. 108–143.

[299] A. Lindenmayer. "Mathematical models for cellular interactions in development I. Filaments with one-sided inputs". In: *Journal of Theoretical Biology* 18.3 (1968), pp. 280–299.

[300] S. Wolfram. "Statistical mechanics of cellular automata". In: *Reviews of Modern Physics* 55.3 (1983), pp. 601–644.

[301] T. Head. "Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors". In: *Bulletin of Mathematical Biology* 49.6 (1987), pp. 737–759.

[302] G. Păun. "Introduction to membrane computing". In: *Applications of Membrane Computing*. Springer, 2006, pp. 1–42.

[303] L. M. Adleman et al. "Molecular computation of solutions to combinatorial problems". In: *Science* 266.5187 (1994), pp. 1021–1024.

[304] P. Siuti, J. Yazbek, and T. K. Lu. "Synthetic circuits integrating logic and memory in living cells". In: *Nature Biotechnology* 31 (2013), pp. 448–452.

[305] L. M. Adleman. "Computing with DNA". In: *Scientific American* 279.8 (1998), pp. 34–41.

[306] V. B. Teif. "Predicting gene-regulation functions: lessons from temperate bacteriophages". In: *Biophysical Journal* 98.7 (2010), pp. 1247–1256.

[307] R. C. Burgess, T. Misteli, and P. Oberdoerffer. "DNA damage, chromatin, and transcription: the trinity of aging". In: *Current Opinion in Cell Biology* 24.6 (2012), pp. 724–730.

[308] V. P. Zediak, E. J. Wherry, and S. L. Berger. "The contribution of epigenetic memory to immunologic memory". In: *Current Opinion in Genetics & Development* 21.2 (2011), pp. 154–159.

[309] F. Celada. "The cellular basis of immunologic memory". In: *Progress in Allergy*. Ed. by P. Kallos, B. Waksman, and A. de Weck. Vol. 15. Basel: Karger, 1971, pp. 223–267.

[310] S. Tonna, A. El-Osta, M. E. Cooper, and C. Tikellis. "Metabolic memory and diabetic nephropathy: potential role for epigenetic mechanisms". In: *Nature Reviews Nephrology* 6.6 (2010), pp. 332–341.

[311] Q. Guan, S. Haroon, D. G. Bravo, J. L. Will, and A. P. Gasch. "Cellular Memory of Acquired Stress Resistance in *Saccharomyces cerevisiae*". In: *Genetics* 192.2 (2012), pp. 495–505.

[312] M. Jaskiewicz, U. Conrath, and C. Peterhänsel. "Chromatin modification acts as a memory for systemic acquired resistance in the plant stress response". In: *EMBO Reports* 12.1 (2010), pp. 50–55.

[313] H. Bolouri and E. H. Davidson. "The gene regulatory network basis of the "community effect," and analysis of a sea urchin embryo example". In: *Developmental Biology* 340.2 (2010), pp. 170–178.

[314] N. Dershowitz. "Computing with rewrite systems". In: *Information and Control* 65.2 (1985), pp. 122–157.

[315] N. Dershowitz and J.-P. Jouannaud. *Rewrite systems*. Citeseer, 1989.

[316] L. Yang, C. Lin, C. Jin, J. C. Yang, B. Tanasa, et al. "lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs". In: *Nature* 500 (2013), pp. 598–602.

[317] M. J. Flynn. "Very high-speed computing systems". In: *Proceedings of the IEEE* 54.12 (1966), pp. 1901–1909.

[318] F. Mueller-Planitz, H. Klinker, and P. B. Becker. "Nucleosome sliding mechanisms: new twists in a looped history". In: *Nature Structural & Molecular Biology* 20 (2013), pp. 1026–1032.

[319] J. Sparsø and S. Furber, eds. *Principles of Asynchronous Circuit Design*. Kluwer Academic Publishers, 2002.

[320] A. Davis and S. M. Nowick. *An introduction to asynchronous circuit design*. Tech. rep. The Encyclopedia of Computer Science and Technology, 1997.

[321] A. Goldbeter, C. Gérard, D. Gonze, J.-C. Leloup, and G. Dupont. "Systems biology of cellular rhythms". In: *FEBS Letters* 586.18 (2012), pp. 2955–2965.

[322] M. Doi, J. Hirayama, and P. Sassone-Corsi. "Circadian regulator CLOCK is a histone acetyltransferase". In: *Cell* 125.3 (2006), pp. 497–508.

[323] S. Masri and P. Sassone-Corsi. "The circadian clock: a framework linking metabolism, epigenetics and neuronal function". In: *Nature Reviews Neuroscience* 4.1 (2012), pp. 69–75.

[324] H. Abelson, T. F. Knight, G. J. Sussman, et al. *Amorphous computing manifesto*. 1996.

[325] S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraula, et al., eds. *Self-organization in biological systems*. Princeton University Press, 2003.

[326] P. Davies. "The epigenome and top-down causation". In: *Interface Focus* 2.1 (2012), pp. 42–48.

[327] M. Mitchell et al. "Computation in Cellular Automata: A Selected Review". In: *Nonstandard Computation* (1996), pp. 95–140.

[328] P. W. Rothemund. "A DNA and restriction enzyme implementation of Turing machines". In: *DNA Based Computers*. Ed. by R. J. Lipton and E. B. Baum. Vol. 6. American Mathematical Society, 1996, pp. 75–120.

[329] A. Nayebi. "Fast matrix multiplication techniques based on the Adleman-Lipton model". In: *arXiv preprint arXiv:0912.0750* (2009).

[330] G. M. Church, Y. Gao, and S. Kosuri. "Next-generation digital information storage in DNA". In: *Science* 337.6102 (2012), pp. 1628–1628.

[331] S. M. Tan-Wong, J. B. Zaugg, J. Camblong, Z. Xu, D. W. Zhang, et al. "Gene loops enhance transcriptional directionality". In: *Science* 338.6107 (2012), pp. 671–675.

[332] I. Whitehouse, O. J. Rando, J. Delrow, and T. Tsukiyama. "Chromatin remodelling at promoters suppresses antisense transcription". In: *Nature* 450.7172 (2007), pp. 1031–1035.

[333] S. I. Walker and P. C. Davies. "The algorithmic origins of life". In: *Journal of The Royal Society Interface* 10.79 (2013), p. 20120869.

[334] F. Santos and W. Dean. "Epigenetic reprogramming during early development in mammals". In: *Reproduction* 127.6 (2004), pp. 643–651.

[335] D. Görlich, S. Artmann, and P. Dittrich. "Cells as semantic systems". In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1810.10 (2011), pp. 914–923.

[336] H. Qi, A. Blanchard, and T. Lu. "Engineered genetic information processing circuits". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 5.3 (2013), pp. 273–287.

[337] Y. S. Jeong, S. Yeo, J. S. Park, K. K. Lee, and Y. K. Kang. "Gradual development of a genome-wide H3-K9 trimethylation pattern in paternally derived pig pronucleus". In: *Developmental Dynamics* 236.6 (2007), pp. 1509–1516.

[338] D. J. Katz, T. M. Edwards, V. Reinke, and W. G. Kelly. "A *C. elegans* LSD1 Demethylase Contributes to Germline Immortality by Reprogramming Epigenetic Memory". In: *Cell* 137.2 (2009), pp. 308–320.

[339] V. Mutskov and G. Felsenfeld. "Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9". In: *The EMBO Journal* 23.1 (2003), pp. 138–149.

[340] C. B. Millar and M. Grunstein. "Genome-wide patterns of histone modifications in yeast". In: *Nature Reviews Molecular Cell Biology* 7.9 (2006), pp. 657–666.

[341] J. Przybilla, J. Galle, and T. Rohlf. "Is adult stem cell aging driven by conflicting modes of chromatin remodeling?" In: *Bioessays* 34.10 (2012), pp. 841–848.

[342] D. Tillo, N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, et al. "High nucleosome occupancy is encoded at human regulatory sequences". In: *PLOS ONE* 5.2 (2010), e9129.

[343] X. Wang, G. Bryant, M. Floer, D. Spagna, and M. Ptashne. "An effect of DNA sequence on nucleosome occupancy and removal". In: *Nature Structural & Molecular Biology* 18.4 (2011), pp. 507–509.

[344] A. Andrews and K. Luger. "Nucleosome structure(s) and stability: variations on a theme". In: *Annual Review of Biophysics* 40 (2011), pp. 99–117.

[345] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht, C. L. Smith, D. Raha, et al. "Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements". In: *Genome Research* 22 (2012), pp. 1735–1747.

[346] L. Cacchiani. "Simulation and Analysis of Chemical Reactions using Stochastic Differential Equations". MA thesis. University of Torino and University of Edinburgh, 2007.

[347] C. Furusawa and K. Kaneko. "Epigenetic Feedback Regulation Accelerates Adaptation and Evolution". In: *PLOS ONE* 8.5 (2013), e61251.

[348] D. T. Gillespie. "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". In: *Journal of Computational Physics* 22 (1976), pp. 403–434.

[349] D. T. Gillespie. "Exact Stochastic Simulation of Coupled Chemical Reactions". In: *The Journal of Physical Chemistry* 81 (1977), pp. 2340–2361.

[350] H. Li and L. Petzold. *Logarithmic direct method for discrete stochastic simulation of chemically reacting systems*. Tech. rep. Santa Barbara, CA: UCSB Computer Science and Engineering Group, 2006.

[351] A. Slepoy, A. P. Thompson, and S. J. Plimpton. "A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks". In: *The Journal of Chemical Physics* 128 (2008), p. 205101.

[352] M. A. Gibson and J. Bruck. "Efficient exact stochastic simulation of chemical systems with many species and many channels". In: *The Journal of Physical Chemistry A* 104.9 (2000), pp. 1876–1889.

[353] D. T. Gillespie. "Approximate accelerated stochastic simulation of chemically reacting systems". In: *The Journal of Chemical Physics* 115.4 (2001), pp. 1716–1733.

[354] T. Tian and K. Burrage. "Binomial leap methods for simulating stochastic chemical kinetics". In: *The Journal of Chemical Physics* 121 (2004), p. 10356.

[355] A. Chatterjee, D. G. Vlachos, and M. A. Katsoulakis. "Binomial distribution based τ-leap accelerated stochastic simulation". In: *The Journal of Chemical Physics* 122 (2005), p. 024112.

[356] Y. Cao, D. T. Gillespie, and L. R. Petzold. "Efficient step size selection for the tau-leaping simulation method". In: *The Journal of Chemical Physics* 124 (2006), p. 044109.

[357] Y. Cao, D. T. Gillespie, and L. R. Petzold. "The slow-scale stochastic simulation algorithm". In: *The Journal of Chemical Physics* 122 (2005), p. 014116.

[358] Y. Cao, D. T. Gillespie, and L. R. Petzold. "Adaptive explicit-implicit tau-leaping method with automatic tau selection". In: *The Journal of Chemical Physics* 126.22 (2007), pp. 224101–224101.

[359] N. Le Novere and T. S. Shimizu. "STOCHSIM: modelling of stochastic biomolecular processes". In: *Bioinformatics* 17.6 (2001), pp. 575–576.

[360] Z. Liu and Y. Cao. "Detailed comparison between StochSim and SSA". In: *IET Systems Biology* 2.5 (2008), pp. 334–341.

[361] C. Jiang and B. F. Pugh. "Nucleosome positioning and gene regulation: advances through genomics". In: *Nature Reviews Genetics* 10.3 (2009), pp. 161–172.

[362] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, et al. "Nucleosome organization in the *Drosophila* genome". In: *Nature* 453.7193 (2008), pp. 358–362.

[363] A. N. Yadon, D. Van de Mark, R. Basom, J. Delrow, I. Whitehouse, et al. "Chromatin remodeling around nucleosome-free regions leads to repression of noncoding RNA transcription". In: *Molecular and Cellular Biology* 30.21 (2010), pp. 5110–5122.

[364] M. D. VerMilyea, L. O'Neill, and B. M. Turner. "Transcription-independent heritability of induced histone modifications in the mouse preimplantation embryo". In: *PLOS ONE* 4.6 (2009), e6086.

[365] M. Eigen. "Selforganization of matter and the evolution of biological macromolecules". In: *Naturwissenschaften* 58.10 (1971), pp. 465–523.

[366] J. A. Hackett, J. J. Zylicz, and M. A. Surani. "Parallel mechanisms of epigenetic reprogramming in the germline". In: *Trends in Genetics* 28.4 (2012), pp. 164–174.

[367] K. Adachi and H. R. Schöler. "Directing reprogramming to pluripotency by transcription factors". In: *Current Opinion in Genetics & Development* 22.5 (2012), pp. 416–422.

[368] A. Watanabe, Y. Yamada, and S. Yamanaka. "Epigenetic regulation in pluripotent stem cells: a key to breaking the epigenetic barrier". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1609 (2013), p. 20120292.

[369] D. Mazza, F. Mueller, T. J. Stasevich, and J. G. McNally. "Convergence of chromatin binding estimates in live cells". In: *Nature Methods* 10.8 (2013), pp. 691–692.

[370] G. Leguizamón and E. Alba. "Ant Colony Based Algorithms for Dynamic Optimization Problems". In: *Metaheuristics for Dynamic Optimization*. Springer, 2013, pp. 189–210.

[371] M. Dorigo and T. Stützle. "Ant colony optimization: overview and recent advances". In: *Handbook of Metaheuristics*. Ed. by M. Gendreau and J.-Y. Potvin. International Series in Operations Research & Management Science. Springer, 2010, pp. 227–263.

[372] D. Tulpan. "Recent patents and challenges on DNA microarray probe design technologies". In: *Recent Patents on DNA and Gene Sequences* 4.3 (2010), pp. 210–217.

[373] C. Lange, L. Mittermayr, J. Dohm, D. Holtgräwe, B. Weisshaar, et al. "High-throughput identification of genetic markers using representational oligonucleotide microarray analysis". In: *Theoretical and Applied Genetics* 121.3 (2010), pp. 549–565.

[374] S. Pascoal, G. Carvalho, O. Vasieva, R. Hughes, A. Cossins, et al. "Transcriptomics and *in vivo* tests reveal novel mechanisms underlying endocrine disruption in an ecological sentinel, Nucella lapillus". In: *Molecular Ecology* 22.6 (2012), pp. 1589–608.

[375] P. Tirumalai and S. Prakash. "Time-dependant gene expression pattern of *Listeria monocytogenes* J0161 in biofilms". In: *Advances in Genomics & Genetics* 2 (2012), pp. 1–18.

[376] K. Sakurai, H. Arai, M. Ishii, and Y. Igarashi. "Transcriptome response to different carbon sources in *Acetobacter aceti*". In: *Microbiology* 157.3 (2011), pp. 899–910.

[377] D. Stekel. *Microarray Bioinformatics*. Cambridge University Press, 2003.

[378] D. Slonim and I. Yanai. "Getting started in gene expression microarray analysis". In: *PLoS Computational Biology* 5.10 (2009), e1000543.

[379] Y. Yang, T. Speed, et al. "Design issues for cDNA microarray experiments". In: *Nature Reviews Genetics* 3.8 (2002), pp. 579–588.

[380] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, et al. "GENCODE: the reference human genome annotation for The ENCODE Project". In: *Genome Research* 22.9 (2012), pp. 1760–1774.

[381] W. J. Kent. "BLAT–the BLAST-like alignment tool". In: *Genome Research* 12.4 (2002), pp. 656–664.

[382] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, et al. "Identification and classification of conserved RNA secondary structures in the human genome". In: *PLoS Computational Biology* 2.4 (2006), e33.

[383] S. Washietl, I. L. Hofacker, and P. F. Stadler. "Fast and reliable prediction of noncoding RNAs". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2454–2459.

[384] I. Hofacker and P. Stadler. "RNAz 2.0: Improved noncoding RNA detection". In: *Pacific Symposium on Biocomputing*. Vol. 15. 2010, pp. 69–79.

[385] J. E. Wilusz, H. Sunwoo, and D. L. Spector. "Long noncoding RNAs: functional surprises from the RNA world". In: *Genes & Development* 23.13 (2009), pp. 1494–1504.

[386] A. Barczak, M. Rodriguez, K. Hanspers, L. Koth, Y. Tai, et al. "Spotted long oligonucleotide arrays for human gene expression analysis". In: *Genome Research* 13.7 (2003), pp. 1775–1785.

[387] C. Chou, C. Chen, T. Lee, and K. Peck. "Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression". In: *Nucleic Acids Research* 32.12 (2004), e99–e99.

[388] R. Russell. "Designing microarray oligonucleotide probes". In: *Briefings in Bioinformatics* 4.4 (2003), pp. 361–367.

[389] S. Lemoine, F. Combes, and S. Le Crom. "An evaluation of custom microarray applications: the oligonucleotide design challenge". In: *Nucleic Acids Research* 37.6 (2009), pp. 1726–1739.

[390] H.-H. Chou, A.-P. Hsia, D. L. Mooney, and P. S. Schnable. "Picky: oligo microarray design for large genomes". In: *Bioinformatics* 20.17 (2004), pp. 2893–2902.

[391] E. Dugat-Bony, M. Missaoui, E. Peyretaillade, C. Biderre-Petit, O. Bouzid, et al. "HiSpOD: probe design for functional DNA microarrays". In: *Bioinformatics* 27.5 (2011), pp. 641–648.

[392] S.-Y. Shin, I.-H. Lee, Y.-M. Cho, K.-A. Yang, and B.-T. Zhang. "EvoOligo: oligonucleotide probe design with multiobjective evolutionary algorithms". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 39.6 (2009), pp. 1606–1616.

[393] W. Li, J. Huang, M. Fan, S. Wang, et al. "MProbe: Computer aided probe design for oligonucleotide microarrays". In: *Applied Bioinformatics* 1 (2002), pp. 163–166.

[394] O. J. Marshall. "PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR". In: *Bioinformatics* 20.15 (2004), pp. 2471–2472.

[395] L. Jourdren, A. Duclos, C. Brion, T. Portnoy, H. Mathis, et al. "Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments". In: *Nucleic Acids Research* 38.10 (2010), e117–e117.

[396] S. Feng and E. R. Tillier. "A fast and flexible approach to oligonucleotide probe design for genomes and gene families". In: *Bioinformatics* 23.10 (2007), pp. 1195–1202.

[397] M. Kane, T. Jatkoe, C. Stumpf, J. Lu, J. Thomas, et al. "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays". In: *Nucleic Acids Research* 28.22 (2000), pp. 4552–4557.

[398] J. Liebich, C. W. Schadt, S. C. Chong, Z. He, S.-K. Rhee, et al. "Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications". In: *Applied and Environmental Microbiology* 72.2 (2006), pp. 1688–1691.

[399] F. M. Giorgi, C. Del Fabbro, and F. Licausi. "Comparative study of RNA-seq-and Microarray-derived coexpression networks in *Arabidopsis thaliana*". In: *Bioinformatics* 29.6 (2013), pp. 717–724.

[400] C. U. Lithner, P. N. Lacor, W.-Q. Zhao, T. Mustafiz, W. L. Klein, et al. "Disruption of neocortical histone H3 homeostasis by soluble Aβ: implications for Alzheimer's disease". In: *Neurobiology of Aging* 34.9 (2013), pp. 2081–2090.

[401] J. Gräff and L.-H. Tsai. "Histone acetylation: molecular mnemonics on the chromatin". In: *Nature Reviews Neuroscience* 14.2 (2013), pp. 97–111.

[402] C. Rouaux, N. Jokic, C. Mbebi, S. Boutillier, J.-P. Loeffler, et al. "Critical loss of CBP/p300 histone acetylase activity by caspase-6 during neurodegeneration". In: *The EMBO Journal* 22.24 (2003), pp. 6537–6549.

[403] A. G. Kazantsev and L. M. Thompson. "Therapeutic application of histone deacetylase inhibitors for central nervous system disorders". In: *Nature Reviews Drug Discovery* 7.10 (2008), pp. 854–868.

[404] R. C. Agis-Balboa, Z. Pavelka, C. Kerimoglu, and A. Fischer. "Loss of HDAC5 Impairs Memory Function: Implications for Alzheimer's Disease". In: *Journal of Alzheimer's Disease* 33.1 (2013), pp. 35–44.

[405] Y. Huang and L. Mucke. "Alzheimer mechanisms and therapeutic strategies". In: *Cell* 148.6 (2012), pp. 1204–1222.

[406] R. A. Boonen, P. van Tijn, and D. Zivkovic. "Wnt signaling in Alzheimer's disease: up or down, that is the question". In: *Ageing Research Reviews* 8.2 (2009), pp. 71–82.

[407] N. C. Inestrosa, C. Montecinos-Oliva, and M. Fuenzalida. "Wnt Signaling: Role in Alzheimer Disease and Schizophrenia". In: *Journal of Neuroimmune Pharmacology* 7.4 (2012), pp. 788–807.

[408] P. Zatta, D. Drago, S. Bolognin, and S. L. Sensi. "Alzheimer's disease, metal ions and metal homeostatic therapy". In: *Trends in Pharmacological Sciences* 30.7 (2009), pp. 346–355.

[409] V. Günther, U. Lindert, and W. Schaffner. "The taste of heavy metals: gene regulation by MTF-1". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1823.9 (2012), pp. 1416–1425.

[410] J. Cooper-Knock, J. Kirby, L. Ferraiuolo, P. R. Heath, M. Rattray, et al. "Gene expression profiling in human neurodegenerative disease". In: *Nature Reviews Neurology* 8 (2012), pp. 518–530.

[411] Y. Yang, E. J. Mufson, and K. Herrup. "Neuronal cell death is preceded by cell cycle events at all stages of Alzheimer's disease". In: *The Journal of Neuroscience* 23.7 (2003), pp. 2557–2563.

[412] C. Moh, J. Z. Kubiak, V. P. Bajic, X. Zhu, M. A. Smith, et al. "Cell cycle deregulation in the neurons of Alzheimer's disease". In: *Cell Cycle in Development*. Springer, 2011, pp. 565–576.

[413] A. Demuro, I. Parker, and G. E. Stutzmann. "Calcium signaling and amyloid toxicity in Alzheimer disease". In: *Journal of Biological Chemistry* 285.17 (2010), pp. 12463–12468.

[414] J.-C. Lambert, B. Grenier-Boley, V. Chouraki, S. Heath, D. Zelenika, et al. "Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis". In: *Journal of Alzheimer's Disease* 20.4 (2010), pp. 1107–1118.

[415] J.-Z. Wang, I. Grundke-Iqbal, and K. Iqbal. "Kinases and phosphatases and tau sites involved in Alzheimer neurofibrillary degeneration". In: *European Journal of Neuroscience* 25.1 (2007), pp. 59–68.

[416] A. C. Paula-Lima, J. Brito-Moreira, and S. T. Ferreira. "Deregulation of excitatory neurotransmission underlying synapse failure in Alzheimer's disease". In: *Journal of Neurochemistry* 126.2 (2013), pp. 191–202.

[417] J. R. Bamburg and G. S. Bloom. "Cytoskeletal pathologies of Alzheimer disease". In: *Cell Motility and the Cytoskeleton* 66.8 (2009), pp. 635–649.

[418] G. Liu, Y. Jiang, P. Wang, R. Feng, N. Jiang, et al. "Cell adhesion molecules contribute to Alzheimer's disease: multiple pathway analyses of two genome-wide association studies". In: *Journal of Neurochemistry* 120.1 (2012), pp. 190–198.

[419] D. W. Wesson, E. Levy, R. A. Nixon, and D. A. Wilson. "Olfactory dysfunction correlates with amyloid-β burden in an Alzheimer's disease mouse model". In: *The Journal of Neuroscience* 30.2 (2010), pp. 505–514.

[420] R. J. Bateman, C. Xiong, T. L. Benzinger, A. M. Fagan, A. Goate, et al. "Clinical and biomarker changes in dominantly inherited Alzheimer's disease". In: *New England Journal of Medicine* 367.9 (2012), pp. 795–804.

[421] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, et al. "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study". In: *The Lancet Neurology* 2.4 (2013), pp. 357–367.

[422] M. Jucker and L. C. Walker. "Self-propagation of pathogenic protein aggregates in neurodegenerative diseases". In: *Nature* 501.7465 (2013), pp. 45–51.

[423] P. Tiraboschi, L. Hansen, L. Thal, and J. Corey-Bloom. "The importance of neuritic plaques and tangles to the development and evolution of AD". In: *Neurology* 62.11 (2004), pp. 1984–1989.

[424] E. Ciarlo, S. Massone, I. Penna, M. Nizzari, A. Gigoni, et al. "An intronic ncRNA-dependent regulation of SORL1 expression affecting Aβ formation is upregulated in post-mortem Alzheimer's disease brain samples". In: *Disease Models & Mechanisms* 6.2 (2012), pp. 424–433.

[425] F. Hernandez and J. Avila. "Tauopathies". In: *Cellular and Molecular Life Sciences* 64.17 (2007), pp. 2219–2233.

[426] R. C. Gentleman, V. J. Carey, D. M. Bates, et al. "Bioconductor: Open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (2004), R80.

[427] M. Reimers and J. N. Weinstein. "Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases". In: *BMC Bioinformatics* 6.1 (2005), p. 166.

[428] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19.2 (2003), pp. 185–193.

[429] H. Nakaya, P. Amaral, R. Louro, A. Lopes, A. Fachel, et al. "Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription". In: *Genome Biology* 8.3 (2007), R43.

[430] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. "GeneCards: integrating information about genes, proteins and diseases". In: *Trends in Genetics* 13.4 (1997), p. 163.

[431] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support." In: *Bioinformatics* 14.8 (1998), pp. 656–664.

[432] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), pp. 25–29.

[433] U. Dreses-Werringloer, J.-C. Lambert, V. Vingtdeux, H. Zhao, H. Vais, et al. "A Polymorphism in CALHM1 Influences Ca_{2+} Homeostasis, Aβ Levels, and Alzheimer's Disease Risk". In: *Cell* 133.7 (2008), pp. 1149–1161.

[434] L. Tapia-Arancibia, E. Aliaga, M. Silhol, and S. Arancibia. "New insights into brain BDNF function in normal aging and Alzheimer disease". In: *Brain Research Reviews* 59.1 (2008), pp. 201–220.

[435] K. Schindowski, K. Belarbi, and L. Buee. "Neurotrophic factors in Alzheimer's disease: role of axonal transport". In: *Genes, Brain and Behavior* 7.s1 (2008), pp. 43–56.

[436] X. Wu, T. Kihara, A. Akaike, T. Niidome, and H. Sugimoto. "PI3K/Akt/mTOR signaling regulates glutamate transporter 1 in astrocytes". In: *Biochemical and Biophysical Research Communications* 393.3 (2010), pp. 514–518.

[437] S. Wang, U. Qaisar, X. Yin, and P. Grammas. "Gene expression profiling in Alzheimer's disease brain microvessels". In: *Journal of Alzheimer's Disease* 31.1 (2012), pp. 193–205.

[438] P. Katsel, W. Tan, P. Fam, D. P. Purohit, and V. Haroutunian. "Cycle Checkpoint Abnormalities during Dementia: A Plausible Association with the Loss of Protection against Oxidative Stress in Alzheimer's Disease". In: *PLOS ONE* 8.7 (2013), e68361.

[439] H. Rhinn, R. Fujita, L. Qiang, R. Cheng, J. H. Lee, et al. "Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease". In: *Nature* 500.7460 (2013), pp. 45–50.

[440] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (2009), pp. 1088–1093.

[441] S. Massone, I. Vassallo, G. Fiorino, M. Castelnuovo, F. Barbieri, et al. "17A, a novel non-coding RNA, regulates GABA B alternative splicing and signaling in response to inflammatory stimuli and in Alzheimer disease". In: *Neurobiology of Disease* 41.2 (2011), pp. 308–317.

[442] M. A. Alarcón, M. A. Medina, Q. Hu, M. E. Avila, B. I. Bustos, et al. "A novel functional low-density lipoprotein receptor-related protein 6 gene alternative splice variant is associated with Alzheimer's disease". In: *Neurobiology of Aging* 34.6 (2012), 1709.e9–1709.e18.

[443] T. Kavanagh, J. D. Mills, W. S. Kim, G. M. Halliday, and M. Janitz. "Pathway Analysis of the Human Brain Transcriptome in Disease". In: *Journal of Molecular Neuroscience* (2012), pp. 1–9.

[444] D. Scheuner, C. Eckman, M. Jensen, X. Song, M. Citron, et al. "Secreted amyloid β–protein similar to that in the senile plaques of Alzheimer's disease is increased *in vivo* by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease". In: *Nature Medicine* 2.8 (1996), pp. 864–870.

[445] M. Talkowski, G. Maussion, L. Crapper, J. Rosenfeld, I. Blumenthal, et al. "Disruption of a Large Intergenic Noncoding RNA in Subjects with Neurodevelopmental Disabilities". In: *The American Journal of Human Genetics* 91.6 (2012), pp. 1128–1134.

[446] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC Bioinformatics* 10.1 (2009), p. 48.

[447] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. "Discovering motifs in ranked lists of DNA sequences". In: *PLoS Computational Biology* 3.3 (2007), e39.

[448] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, et al. "GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (2004), pp. 3710–3715.

[449] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc. "REVIGO summarizes and visualizes long lists of gene ontology terms". In: *PLOS ONE* 6.7 (2011), e21800.

[450] N. N. Karpova. "Role of BDNF epigenetics in activity-dependent neuronal plasticity". In: *Neuropharmacology* (2013).

[451] J. L. Stein, S. E. Medland, A. A. Vasquez, D. P. Hibar, R. E. Senstad, et al. "Identification of common variants associated with human hippocampal and intracranial volumes". In: *Nature Genetics* 44.5 (2012), pp. 552–561.

[452] G. Whelan, E. Kreidl, G. Wutz, A. Egner, J.-M. Peters, et al. "Cohesin acetyltransferase Esco2 is a cell viability factor and is required for cohesion in pericentric heterochromatin". In: *The EMBO Journal* 31.1 (2011), pp. 71–82.

[453] T. Alenghat, K. Meyers, S. E. Mullican, K. Leitner, A. Adeniji-Adele, et al. "Nuclear receptor corepressor and histone deacetylase 3 govern circadian metabolic physiology". In: *Nature* 456.7224 (2008), pp. 997–1000.

[454] M. Nishiyama, A. I. Skoultchi, and K. I. Nakayama. "Histone H1 Recruitment by CHD8 Is Essential for Suppression of the Wnt–β-Catenin Signaling Pathway". In: *Molecular and Cellular Biology* 32.2 (2012), pp. 501–512.

[455] M.-A. Hakimi, D. A. Bochar, J. A. Schmiesing, Y. Dong, O. G. Barak, et al. "A chromatin remodelling complex that loads cohesin onto human chromosomes". In: *Nature* 418.6901 (2002), pp. 994–998.

[456] H.-g. Lee, G. Casadesus, X. Zhu, R. J. Castellani, A. McShea, et al. "Cell cycle re-entry mediated neurodegeneration and its treatment role in the pathogenesis of Alzheimer's disease". In: *Neurochemistry International* 54.2 (2009), pp. 84–88.

[457] D. J. Bonda, H. Lee, W. Kudo, X. Zhu, M. A. Smith, et al. "Pathological implications of cell cycle re-entry in Alzheimer disease". In: *Expert Reviews in Molecular Medicine* 12 (2010), e19.

[458] A. McShea, H.-g. Lee, R. B. Petersen, G. Casadesus, I. Vincent, et al. "Neuronal cell cycle re-entry mediates Alzheimer disease-type changes". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1772.4 (2007), pp. 467–472.

[459] O. Ogawa, X. Zhu, H.-G. Lee, A. Raina, M. E. Obrenovich, et al. "Ectopic localization of phosphorylated histone H3 in Alzheimer's disease: a mitotic catastrophe?" In: *Acta Neuropathologica* 105.5 (2003), pp. 524–528.

[460] S. Washietl, S. Findeiß, S. A. Müller, S. Kalkhof, M. von Bergen, et al. "RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data". In: *RNA* 17.4 (2011), pp. 578–594.

[461] W. Li, W. Yang, and X.-J. Wang. "Pseudogenes: pseudo or real functional elements?" In: *Journal of Genetics and Genomics* 40.4 (2013), pp. 171–177.

[462] B. Pei, C. Sisu, A. Frankish, C. Howald, L. Habegger, et al. "The GENCODE pseudogene resource". In: *Genome Biology* 13 (2012), R51.

[463] Y.-Z. Wen, L.-L. Zheng, L.-H. Qu, F. J. Ayala, and Z.-R. Lun. "Pseudogenes are not pseudo any more". In: *RNA Biology* 9.1 (2012), pp. 27–32.

[464] R. C. Pink and D. R. Carter. "Pseudogenes as regulators of biological function". In: *Essays in Biochemistry* 54.1 (2013), pp. 103–112.

[465] K. Herrup. "The contributions of unscheduled neuronal cell cycle events to the death of neurons in Alzheimer's disease." In: *Frontiers in Bioscience (Elite Edition)* 4 (2012), p. 2101.

[466] J. Kukucka, T. Wyllie, J. Read, L. Mahoney, and C. Suphioglu. "Human neuronal cells: epigenetic aspects". In: *BioMolecular Concepts* 4.4 (2013), pp. 319–333.

[467] M. Guttman, J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, et al. "lincRNAs act in the circuitry controlling pluripotency and differentiation". In: *Nature* 477.7364 (2011), pp. 295–300.

[468] K. Liu, Z. Yan, Y. Li, and Z. Sun. "Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis". In: *Bioinformatics* 29.17 (2013), pp. 2221–2222.

[469] J. Zhu, M. Adli, J. Zou, G. Verstappen, M. Coyne, et al. "Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues". In: *Cell* 152.3 (2013), pp. 642–654.

[470] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, et al. "The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements". In: *Nature Biotechnology* 24.9 (2006), pp. 1151–1161.

[471] G. T. Sutherland, M. Janitz, and J. J. Kril. "Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics?" In: *Journal of Neurochemistry* 116.6 (2011), pp. 937–946.

[472] P. Katsel, W. Tan, and V. Haroutunian. "Gain in brain immunity in the oldest-old differentiates cognitively normal from demented individuals". In: *PLOS ONE* 4.10 (2009), e7642.

[473] P. Katsel, C. Li, and V. Haroutunian. "Gene expression alterations in the sphingolipid metabolism pathways during progression of dementia and Alzheimer's disease: a shift toward ceramide accumulation at the earliest recognizable stages of Alzheimer's disease?" In: *Neurochemical Research* 32.4-5 (2007), pp. 845–856.

[474] A. Koppelkamm, B. Vennemann, S. Lutz-Bonengel, T. Fracasso, and M. Vennemann. "RNA integrity in post-mortem samples: influencing parameters and implications on RT-qPCR assays". In: *International Journal of Legal Medicine* 125.4 (2011), pp. 573–580.

[475] H. Braak and E. Braak. "Neuropathological stageing of Alzheimer-related changes". In: *Acta Neuropathologica* 82.4 (1991), pp. 239–259.

[476] K. Bossers, K. T. Wirz, G. F. Meerhoff, A. H. Essing, J. W. van Dongen, et al. "Concerted changes in transcripts in the prefrontal cortex precede neuropathology in Alzheimer's disease". In: *Brain* 133.12 (2010), pp. 3699–3723.

[477] V. Haroutunian, P. Katsel, and J. Schmeidler. "Transcriptional vulnerability of brain regions in Alzheimer's disease and dementia". In: *Neurobiology of Aging* 30.4 (2009), pp. 561–573.

[478] N. C. Berchtold, D. H. Cribbs, P. D. Coleman, J. Rogers, E. Head, et al. "Gene expression changes in the course of normal brain aging are sexually dimorphic". In: *Proceedings of the National Academy of Sciences* 105.40 (2008), pp. 15605–15610.

[479] N. E. Buchler and M. Louis. "Molecular titration and ultrasensitivity in regulatory networks". In: *Journal of Molecular Biology* 384.5 (2008), pp. 1106–1119.

[480] T. A. Ishunina and D. F. Swaab. "Decreased alternative splicing of estrogen receptor-α mRNA in the Alzheimer's disease brain". In: *Neurobiology of Aging* 33.2 (2012), pp. 286–296.

[481] M. J. Chiocco, X. Zhu, D. Walther, O. Pletnikova, J. C. Troncoso, et al. "Fine Mapping of Calcineurin (PPP3CA) Gene Reveals Novel Alternative Splicing Patterns, Association of 5'UTR Trinucleotide Repeat With Addiction Vulnerability, and Differential Isoform Expression in Alzheimer's Disease". In: *Substance Use & Misuse* 45.11 (2010), pp. 1809–1826.

[482] J. Shi, W. Qian, X. Yin, K. Iqbal, I. Grundke-Iqbal, et al. "Cyclic AMP-dependent Protein Kinase Regulates the Alternative Splicing of Tau Exon 10: A mechanism involved in tau pathology of Alzheimer disease". In: *Journal of Biological Chemistry* 286.16 (2011), pp. 14639–14648.

[483] J. Merkin, C. Russell, P. Chen, and C. Burge. "Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues". In: *Science* 338.6114 (2012), pp. 1593–1599.

[484] N. Barbosa-Morais, M. Irimia, Q. Pan, H. Xiong, S. Gueroussov, et al. "The Evolutionary Landscape of Alternative Splicing in Vertebrate Species". In: *Science* 338.6114 (2012), pp. 1587–1593.

[485] J. B. Anderson and R. Johnnesson. *Understanding information transmission*. Vol. 18. Wiley-IEEE Press, 2006.

[486] S. P. Khare, F. Habib, R. Sharma, N. Gadewal, S. Gupta, et al. "HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes". In: *Nucleic Acids Research* 40.D1 (2012), pp. D337–D342.

[487] V. Migliori, J. Müller, S. Phalke, D. Low, M. Bezzi, et al. "Symmetric dimethylation of H3R2 is a newly identified histone mark that supports euchromatin maintenance". In: *Nature Structural & Molecular Biology* 19.2 (2012), pp. 136–144.

[488] A. P. Jack, S. Bussemer, M. Hahn, S. Pünzeler, M. Snyder, et al. "H3K56me3 Is a Novel, Conserved Heterochromatic Mark That Largely but Not Completely Overlaps with H3K9me3 in Both Regulation and Localization". In: *PLOS ONE* 8.2 (2013), e51765.

[489] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2012.

[490] M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, et al. "`MITOS`: Improved *de novo* Metazoan Mitochondrial Genome Annotation". In: *Molecular Phylogenetics and Evolution* 69.2 (2012), pp. 313–319.

[491] D. Bu, K. Yu, S. Sun, C. Xie, G. SkogerbØ, et al. "NONCODE v3.0: integrative annotation of long noncoding RNAs". In: *Nucleic Acids Research* 40.Database issue (2012), pp. D210–D215.

[492] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick. "lncRNAdb: a reference database for long noncoding RNAs". In: *Nucleic Acids Research* 39.Database issue (2011), pp. D146–D151.

[493] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, et al. "fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences". In: *Nucleic Acids Research* 35.Database issue (2007), pp. D145–D148.

[494] K. C. Pang, S. Stephen, M. E. Dinger, P. G. Engström, B. Lenhard, et al. "RNAdb 2.0–an expanded database of mammalian non-coding RNAs". In: *Nucleic Acids Research* 35.Database issue (2007), pp. D178–D182.

[495] C. Yamasaki, K. Murakami, Y. Fujii, Y. Sato, E. Harada, et al. "The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts". In: *Nucleic Acids Research* 36.Database issue (2008), p. D793.

[496] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy". In: *Nucleic Acids Research* 40.Database issue (2012), pp. D130–D135.

[497] S. Bartschat, S. Kehr, H. Tafer, P. Stadler, and J. Hertel. "`snoStrip`: A snoRNA annotation pipeline". in press. 2013.

[498] S. Guil and M. Esteller. "*Cis*-acting noncoding RNAs: friends and foes". In: *Nature Structural & Molecular Biology* 19.11 (2012), pp. 1068–1075.

[499] A. Relógio, C. Schwager, A. Richter, W. Ansorge, and J. Valcárcel. "Optimization of oligonucleotide-based DNA microarrays". In: *Nucleic Acids Research* 30.11 (2002), e51–e51.

[500] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.23 (2009), pp. 9362–9367.

[501] G. K. Smyth et al. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), p. 3.

[502] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300.

[503] J. D. Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[504] K. Strimmer. "A unified approach to false discovery rate estimation". In: *BMC Bioinformatics* 9.1 (2008), p. 303.

[505] K. Strimmer. "fdrtool: a versatile R package for estimating local and tail area-based false discovery rates". In: *Bioinformatics* 24.12 (2008), pp. 1461–1462.

[506] L. Bertram, M. McQueen, K. Mullin, D. Blacker, and R. Tanzi. *The AlzGene Database. Alzheimer Research Forum*. 2006. URL: http://www.alzgene.org (Last accessed: September 20, 2013).

[507] M. G. Tan, W.-T. Chua, M. M. Esiri, A. D. Smith, H. V. Vinters, et al. "Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease". In: *Journal of Neuroscience Research* 88.6 (2010), pp. 1157–1169.

[508] M. G. Ravetti, O. A. Rosso, R. Berretta, and P. Moscato. "Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus' gene expression profiles in Alzheimer's disease". In: *PLOS ONE* 5.4 (2010), e10153.

[509] V. Colangelo, J. Schurr, M. J. Ball, R. P. Pelaez, N. G. Bazan, et al. "Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: Transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling". In: *Journal of Neuroscience Research* 70.3 (2002), pp. 462–473.

[510] S. D. Ginsberg, M. J. Alldred, and S. Che. "Gene expression levels assessed by CA1 pyramidal neuron and regional hippocampal dissections in Alzheimer's disease". In: *Neurobiology of Disease* 45.1 (2012), pp. 99–107.

[511] G. Yeo and C. B. Burge. "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals". In: *Journal of Computational Biology* 11.2-3 (2004), pp. 377–394.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 4. November 2013

Christian Arnold